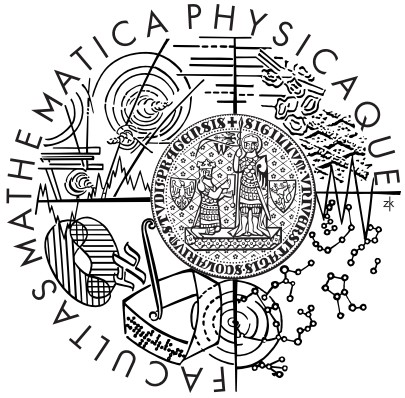
Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Stanislav Nagy

Hloubka funkcionálních dat

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. RNDr. Daniel Hlubinka, Ph.D.

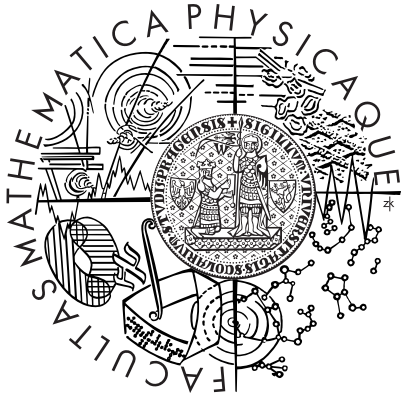
Studijní program: Matematika
Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie

Praha 2011

Univerzita Karlova v Prahe

Matematicko-fyzikálna fakulta

DIPLOMOVÁ PRÁCA



Stanislav Nagy

Hĺbka funkcionálnych dát

Katedra pravdepodobnosti a matematickej štatistiky

Vedúci diplomovej práce: doc. RNDr. Daniel Hlubinka, Ph.D.

Študijný program: Matematika
Študijný obor: Pravdepodobnosť, matematická štatistika a ekonometria

Praha 2011

Chcem poďakovať v prvom rade doc. RNDr. Danielovi Hlubinkovi, Ph.D. za skvelé vedenie a veľa výborných nápadov, prof. RNDr. Janovi Malému, DrSc. za ochotu a rady ohľadom používania matematickej analýzy, Mgr. Zdeňkovi Hlávkovi, Ph.D. za pomoc pri technických problémoch, Ing. Marekovi Omelkovi, Ph.D. za poskytnutie literatúry, prof. RNDr. Jaromírovi Antochovi, CSc. a prof. RNDr. Jane Jurečkovej, DrSc. za záujem a cenné rady. Moja vďaka tiež patrí Veronike Stankovianskej za veľkú pomoc s niektorými prekladmi do angličitny, Pavlovi Křížovi, Jirkovi Francovi a Karlovi Lavičkovi záujem a užitočný kritický náhľad na vec.

Ďalej chcem poďakovať celej rodine za podporu a v neposlednej rade tiež Nike Králikovej a Janke Tormovej za to, že sa ma po celú dobu písania snažili udržať pri zmysloch. Bez nich všetkých by sa práca tak ako je nikdy nepodarila.

Prehlasujem, že som svoju diplomovú prácu napísal samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, najmä skutočnosť, že Univerzita Karlova v Prahe má právo na uzavretie licenčnej zmluvy o použití tejto práce ako školského diela podľa § 60 odst. 1 autorského zákona.

V Prahe dňa Stanislav Nagy

Názov práce: Hĺbka funkcionálnych dát
Autor: Stanislav Nagy
Katedra: Katedra pravdepodobnosti a matematickej štatistiky
Vedúci diplomovej práce: doc. RNDr. Daniel Hlubinka, Ph.D.
e-mail vedúceho: hlubinka@karlin.mff.cuni.cz

Abstrakt: Hĺbková funkcia (resp. funkcionál) je moderný neparametrický nástroj štatistickej analýzy (konečnorozmerných) dát s množstvom praktických aplikácií. V práci sa zameriame na možnosti rozšírenia konceptu hĺbky na prípad funkcionálnych dát. V prípade konečnorozmerných funkcionálnych dát využijeme izomorfizmus priestoru funkcií a konečnorozmerného euklidovského priestoru, čo nám umožní zaviesť indukované hĺbky funkcionálnych dát. Dokážeme tvrdenie o vlastnostiach indukovaných hĺbok a na príkladoch si ukážeme možnosti a obmedzenia ich praktického použitia. Ďalej popíšeme a na jednoduchých príkladoch ukážeme výhody aj nevýhody zavedených hĺbkových funkcionálov používaných v literatúre (Fraimanových-Munizovej hĺbok a pásových hĺbok). Na odstránenie najväčšej vyvstávajúcej nevýhody známych hĺbok pre funkcionálne dáta zavedieme novú, K-pásovú hĺbku založenú na rozšírení inferencie zo spojitých na hladké funkcie. Odvodíme niekoľko dôležitých vlastností a na záverečnej simulačnej štúdií ukážeme na príklade riadenej klasifikácie funkcionálnych dát praktickú výhodnosť nového prístupu oproti predchádzajúcim. Na záver porovnáme výpočetnú náročnosť všetkých predstavených hĺbkových funkcionálov.

Kľúčové slová: hĺbka dát, funkcionálne dáta, klasifikácia dát

Title: The Depth of Functional Data
Author: Stanislav Nagy
Department: Department of Probability and Mathematical Statistics
Supervisor: doc. RNDr. Daniel Hlubinka, Ph.D.
Supervisor's e-mail address: hlubinka@karlin.mff.cuni.cz

Abstract: The depth function (functional) is a modern nonparametric statistical analysis tool for (finite-dimensional) data with lots of practical applications. In the present work we focus on the possibilities of the extension of the depth concept onto a functional data case. In the case of finite-dimensional functional data the isomorphism between the functional space and the finite-dimensional Euclidean space will be utilized in order to introduce the induced functional data depths. A theorem about induced depths' properties will be proven and on several examples the possibilities and restraints of it's practical applications will be shown. Moreover, we describe and demonstrate the advantages and disadvantages of the established depth functionals used in the literature (Fraiman-Muniz depths and band depths). In order to facilitate the outcoming drawbacks of known depths, we propose new, K-band depth based on the inference extension from continuous to smooth functions. Several important properties of the K-band depth will be derived. On a final supervised classification simulation study the reasonability of practical use of the new approach will be shown. As a conclusion, the computational complexity of all presented depth functionals will be compared.

Keywords: data depth, functional data, data classification

Obsah

1 Úvod **1**
1.1 Hĺbka dát 1
1.2 Funkcionálne dáta 9

2 Funkcionálne hĺbky **12**
2.1 Indukované hĺbky 12

3 Geometrické hĺbky **16**
3.1 Fraimanova-Munizovej hĺbka 16
3.2 Pásové hĺbky pre konečnorozmerné dáta 22
3.3 Pásové hĺbky pre funkcionálne dáta 31

4 Geometricko-funkcionálna hĺbka **37**
4.1 K-pásová hĺbka 38

5 Klasifikácia funkcionálnych dát **53**
5.1 Porovnanie na základe simulácií - model posunutia v polohe 55
5.2 Porovnanie na základe simulácií - modely posunutia v tvare 58
5.3 Porovnanie na skutočných dátach - rast detí 62

6 Poznámky k výpočetnej náročnosti a záver **66**

A `C++` zdrojové kódy **72**

B `R` implementácia zdrojových kódov **77**

Kapitola 1

Úvod

Problémom, ktorý budeme riešiť, je možnosť zavedenia štatistickej hĺbkovej funkcie pre funkcionálne dáta. Pripomeňme teda v úvode, čo pod štatistickou hĺbkovou funkciou a funkcionálnymi dátami rozumieme.

1.1 Hĺbka dát

Dôležitým nástrojom štatistickej analýzy dát sú v jednorozmernom prípade poradové štatistiky a poradia pozorovaní v náhodnom výbere. Tieto sa využívajú v aplikáciách ako sú neparametrické testy hypotéz alebo robustné odhady parametrov (L-štatistiky).

Poradie je však v jednorozmernom prípade určené prirodzeným usporiadaním bodov na reálnej osi, čoho analógiu pre viacrozmerné prípady už nie je možné odvodiť. Namiesto takéhoto usporiadania sa pre rozdelenie pravdepodobnosti P na nejakom všeobecnom priestore M zavádza *štatistická hĺbková funkcia*

$$D(.;P) : M \to [0,\infty).$$

V prípade, že bude z kontextu zrejmé, voči akému rozdeleniu pravdepodobnosti sa štatistická hĺbková funkcia počíta, budeme zjednodušovať značenie na $D(.)$ namiesto $D(.;P)$. Výlučne však budeme pracovať s obmedzenými a znormovanými hĺbkovými funkciami tvaru

$$D(.;P) : M \to [0,1]. \tag{1.1}$$

Serfling [24] podobne ako Liu et al. [12] popisujú množstvo možností použitia hĺbkovej funkcie na účely analýzy konečnorozmerných dát. Vytknime niekoľko najzaujímavejších.

- **Usporiadanie a poradové štatistiky.** Hĺbková funkcia poskytuje usporiadanie od „centra“ rozdelenia smerom k okrajom vzhľadom k rozdeleniu pravdepodobnosti P. Tým je možné zovšeobecniť pojem poradia a poradových štatistík pre viacrozmerné dáta. Na základe takéhoto usporiadania dát podľa rastúcej hĺbky je možné zaviesť niektoré neparametrické testy založené na usporiadaní (napr. test o strede symetrie rozdelenia pravdepodobnosti ako uvádza Liu [11], alebo neparametrické rozhodovacie pravidlá pre klasifikačnú úlohu ako v ich článkoch popísali Cuevas et al. [4] alebo López-Pintado a Romo [14, 16], pozri kapitolu 5). Úplné usporiadanie viacrozmerných dát ďalej umožňuje zavedenie analógie L-odhadov pre takéto dáta (napr. Fraiman a Meloche [9]).

- **Centrum a centrálne oblasti rozdelenia.** Je možné zaviesť α-*oblasť hĺbky* ako množinu takých bodov $x \in M$, pre ktoré platí

$$D(x;P) > \alpha,$$

teda takých bodov, v ktorých hĺbka nadobúda väčšiu hodnotu ako daná konštanta α. Podobne je možné zaviesť *p-centrálnu oblasť* $C_P(p)$ rozdelenia P pre $p \in (0,1)$ ako α-oblasť hĺbky pre α také, že platí

$$P(x \in M : D(x;P) > \alpha) = p.$$

- **Funkcie tvaru rozdelenia.** *Funkciu mierky* v_P môžeme pomocou p-centrálnych oblastí definovať ako

$$v_P(p) = \lambda(C_P(p)),$$

kde λ označuje nejakú (v prípade $M = \mathbb{R}^d$ Lebesgueovu) mieru na priestore M. Ďalej je možné pomocou transformácií funkcie mierky zaviesť rôzne funkcionály šikmosti a špicatosti pre viacrozmerné dáta. Takého zobrazenia poskytujú informáciu o tvare rozdelenia aj v prípade, že jeho grafické znázornenie je zložité a neprehľadné (najmä v prípade vysokej dimenzionality priestoru M).

- **Medián a kvantily.** Na základe hĺbky je možné do istej miery zovšeobecniť pojmy ako kvantily alebo medián pre mnohorozmerné dáta, napríklad mnohorozmerný medián je možné definovať práve ako bod (resp. množinu bodov) s najväčšou hĺbkou vo výberovom priestore.

Kvantily je možné zovšeobecniť na *kontúry* rozdelenia pravdepodobnosti. Ak ich označíme ako $\mathrm{Cont}_P(p)$, potom ich definujeme ako množiny bodov v priestore M dané predpisom

$$\mathrm{Cont}_P(p) = \partial(C_P(p)),$$

kde $\partial(A)$ označuje hranicu množiny $A \subset M$. Preto môžeme pozorovania, ktoré neležia v (napríklad) 0.95, alebo 0.99-centrálnej oblasti rozdelenia považovať za odľahlé pozorovania voči náhodnému výberu P. S tým je úzko spätá *funkcia odľahlosti* $O(.;P)$, ktorá meria nakoľko je bod z množiny M odľahlým pozorovaním v náhodnom výbere z rozdelenia P. Zavádzame ju pomocou predpisu

$$D(x;P) = \frac{1}{1+O(x;P)}.$$

Pre konečnorozmerné dáta bolo zavedených niekoľko rôznych hĺbkových funkcií. Pretože však budeme v ďalšom niekoľko z nich používať, pripomeňme explicitne definíciu tých najviac používaných a skúmaných.

Pre množinu M budeme v celom ďalšom texte ako $\mathcal{P}(M)$ označovať triedu všetkých pravdepodobnostných rozdelení na borelovskej σ-algebre množiny M.

Najstaršiu a najznámejšiu *polopriestorovu hĺbku* predstavil Tukey [28]. Je založená na koncepcii polopriestoru ako množiny bodov v nie nutne konečnorozmernom vektorovom priestore. Definujme preto najprv polopriestor vo všeobecnom vektorovom priestore.

Definícia:(Polopriestor)
Nech M je vektorový priestor a $L : M \to \mathbb{R}$ lineárne zobrazenie. Potom *(uzavretý) polopriestor* $H \equiv H_L \subset M$ je množina

$$H = \{x \in M : L(x) \geq 0\}. \tag{1.2}$$

Pre bod $x \in \mathbb{R}^d$ je polopriestorová hĺbka tohto bodu voči pravdepodobnostnému rozdeleniu P definovaná ako infimum pravdepodobností všetkých uzavretých polopriestorov obsahujúcich bod x.

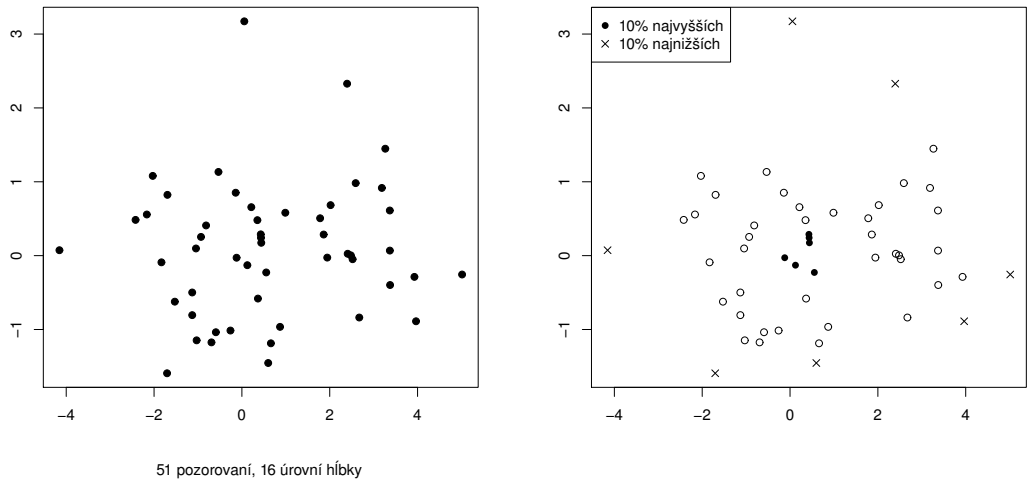
Definícia:(Polopriestorová hĺbka)
Nech M je vektorový priestor, $x \in M$ a $P \in \mathcal{P}(M)$. Nech $\mathcal{H}(M)$ je trieda všetkých uzavretých polopriestorov priestoru M. Potom *polopriestorová hĺbka bodu x vzhľadom k rozdeleniu P* je

$$HD(x;P) = \inf_{H \in \mathcal{H}, x \in H} P(H). \tag{1.3}$$

V prípade náhodného výberu z rozdelenia P počítame výberovú polopriestorovú hĺbku bodu x jednoducho ako $HD\left(x;\hat{P}\right)$, kde $\hat{P}$ označuje empirickú distribúciu náhodného výberu z rozdelenia P.

Ilustrujme použitie polopriestorovej hĺbky na dvoch jednoduchých príkladoch náhodných výberov z dvojrozmerného normálneho rozdelenia.

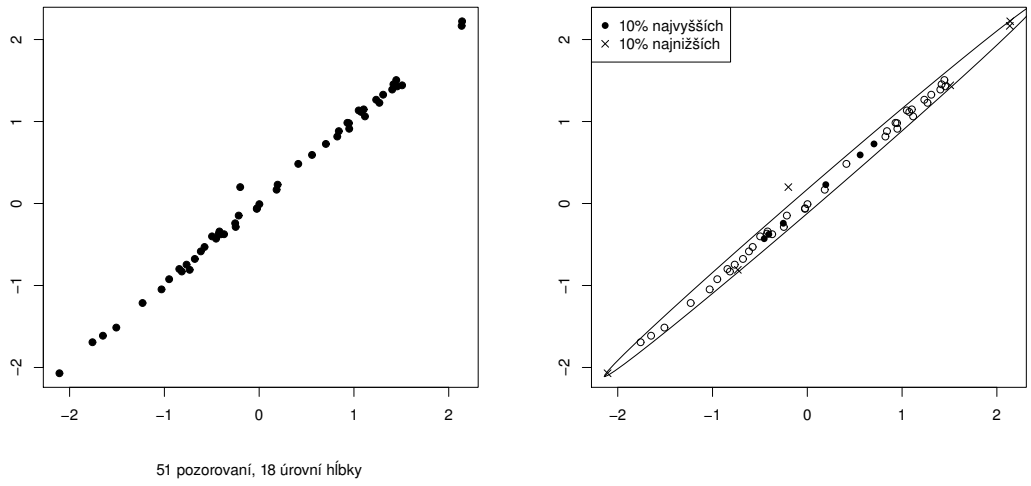
Príklad 1. Prvý náhodný výber má nezávislé zložky a jeho rozsah je 51 bodov. V prvej časti obrázku 1.1 vidíme náhodný výber, v druhej časti je zvýraznených 10 % pozorovaní s najväčšou a 10 % s najmenšou hodnotou polopriestorovej hĺbky voči tomuto náhodnému výberu. Vidíme, že pozorovania s vysokou hodnotou hĺbky sa nachádzajú „uprostred“ zhluku pozorovaní, zatiaľ čo pozorovania s malou hodnotou polopriestorovej hĺbky nájdeme „na okrajoch“ tak, ako by sme to očakávali. Dáta boli rozdelené do 16 úrovní, čo je spôsobené najmä tým, že pomerne veľká časť pozorovaní na okrajoch dostala rovnakú nulovú hĺbku.



Obr. 1.1: Polopriestorová hĺbka a dvojrozmerné normálne rozdelenie.

Príklad 2. Uvažujme ďalej druhý, kontaminovaný náhodný výber 50 bodov z dvojrozmerného normálneho rozdelenia so silne korelovanými zložkami X a Y, to znamená $X \approx aY + b$ pre nejaké konštanty $a, b \in \mathbb{R}$. Kontaminácia je jednobodová a spočíva v tom, že do náhodného výberu sme umelo vniesli chybné pozorovanie ležiace mimo 99% konfidenčnej množiny, ale (v rámci euklidovskej metriky) blízko strednej hodnoty rozdelenia náhodného vektoru $(X,Y)^T$. Preto môžeme kontaminujúce pozorovanie považovať za silne odľahlé.

Počítajme polopriestorovú hĺbku bodov takéhoto náhodného výberu a špeciálne kontaminujúceho bodu. Ako vidíme na obrázku 1.2, kontaminujúci bod ležiaci mimo konfidenčnej elipsy bol spoľahlivo odhalený ako odľahlý a dostal nulovú hĺbku. Za pozorovania typické voči náhodnému výberu boli správne označené body ležiace blízko strednej hodnoty rozdelenia neporušujúce variančnú štruktúru. Medzi pozorovaniami s malou hodnotou hĺbky však nájdeme aj také, ktoré sa voľným okom nedajú identifikovať ako extrémne alebo odľahlé. Výrazne totiž neporušujú štruktúru náhodného výberu a nedajú sa ani označiť za extrémne v žiadnej zložke. Polopriestorová hĺbka ich však kvôli kombinácii toho, že sa trochu vychyľujú zo štruktúry a neležia úplne v centre rozdelenia označuje za kandidátov podozrivých z odľahlosti.



Obr. 1.2: Polopriestorová hĺbka a kontaminované dvojrozmerné normálne rozdelenie.

Zaujímavou vlastnosťou polopriestorovej hĺbky je, že aj keď bola pôvodne definovaná iba v konečnorozmerných euklidovských priestoroch $\mathbb{R}^d$, vďaka všeobecnej definícii polopriestoru 1.2 sme ju v 1.3 mohli prirodzene rozšíriť na prípad nie nutne konečnorozmerného vektorového priestoru. Triviálne rozšírenie polopriestorovej hĺbky až na funkcionálny prípad však nie je dobrým riešením nášho problému. Vo vysokorozmernom a špeciálne v nekonečnorozmernom prípade má totiž polopriestorová hĺbka niektoré zvláštne vlastnosti na ktoré bolo poukázal Chaudhuri [3] a stáva sa nepoužiteľnou. Preto sa takýmto rozšírením polopriestorovej hĺbky na funkcionálne dáta v práci ďalej nebudeme zaoberať.

Ďalšia pre nás zaujímavá hĺbková funkcia je *simplexová hĺbka*, ktorú definovala a vlastnosti skúmala Liu [11]. Simplexová hĺbka meria pravdepodobnosť, že bod $x \in \mathbb{R}^d$ bude ležať v d-rozmernom náhodnom simplexe tvorenom náhodným výberom $(d+1)$ pozorovaní z rozdelenia P. Označujme ďalej pre lineárne nezávislú množinu

bodov $\{x_i\}_{i=1}^{d+1} \subset \mathbb{R}^d$ simplex tvorený týmito bodmi ako $\mathbb{S}_{x_1,\ldots,x_{d+1}}$. V prípade, že táto množina bodov bude lineárne závislá, označujme rovnakým spôsobom konvexný obal týchto bodov.

Definícia:(Simplexová hĺbka-populačná verzia)
Nech $x \in \mathbb{R}^d$ a $P \in \mathcal{P}\left(\mathbb{R}^d\right)$. Nech $X_1,\ldots,X_{d+1}$ je náhodný výber z rozdelenia P. Potom *simplexová hĺbka bodu x vzhľadom k rozdeleniu P* je

$$SD(x;P) = P\left(x \in \mathbb{S}_{X_1,\ldots,X_{d+1}}\right). \tag{1.4}$$

Výberová verzia simplexovej hĺbky dát je založená na odhade pravdepodobnosti 1.4 pomocou U-štatistiky. Pripomeňme teda najprv, čo budeme pod pojmom U-štatistiky rozumieť.

Definícia:(U-štatistika)
Nech $j \in \mathbb{N}$, $h : M^j \to \mathbb{R}$ je merateľné zobrazenie a nech $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia $P \in \mathcal{P}(M)$. Potom štatistika tvaru

$$T(\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\le i_1<\ldots<i_j\le n} h\left(X_{i_1},\ldots,X_{i_j}\right)$$

je *U-štatistika rádu j s jadrom h.*

Zaveďme teda výberovú verziu simplexovej hĺbky.

Definícia:(Simplexová hĺbka-výberová verzia)
Nech $x \in \mathbb{R}^d$ a $P \in \mathcal{P}\left(\mathbb{R}^d\right)$. Nech $n \ge d+1$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Potom *simplexová hĺbka bodu x vzhľadom k náhodnému výberu* $\mathbb{X}$ je

$$SD_n(x;\mathbb{X}) = \binom{n}{d+1}^{-1} \sum_{1\le i_1<\ldots<i_{d+1}\le n} \mathbb{I}\left[x \in \mathbb{S}_{X_{i_1},\ldots,X_{i_{d+1}}}\right],$$

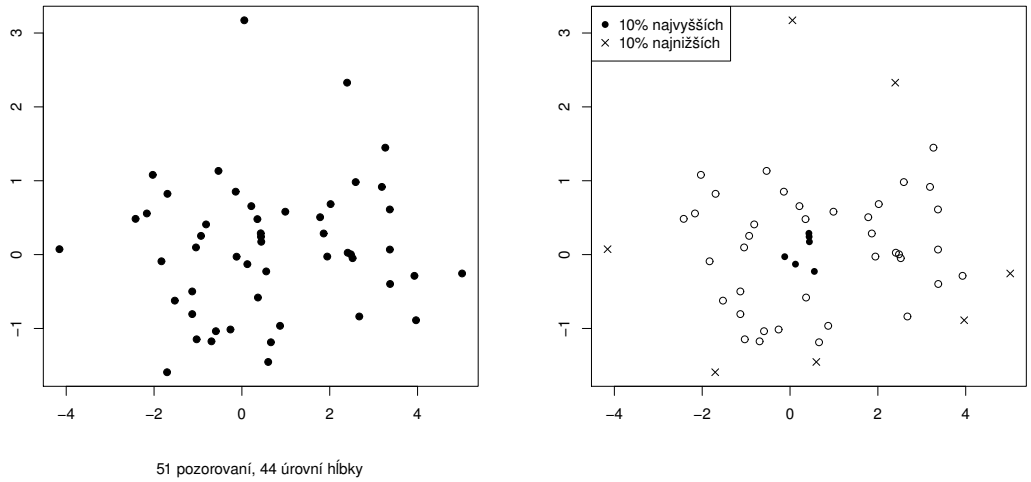
kde $\mathbb{I}[A]$ označuje indikátor javu A.

Použime simplexovú hĺbku na dáta z príkladov 1 a 2 pre porovnanie s polopriestorovou hĺbkou.

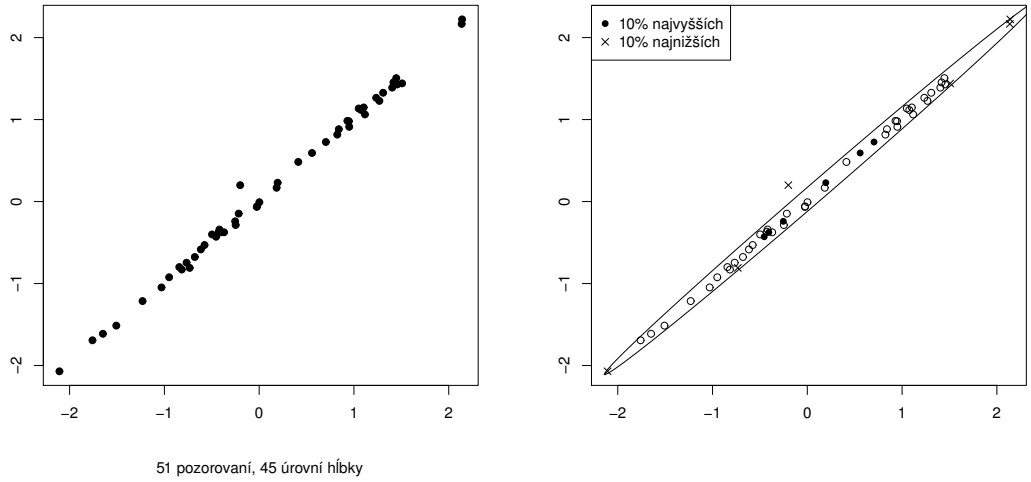
Príklad 3. Ak na analýzu náhodného výberu z dvojrozmerného normálneho rozdelenia z príkladu 1 použijeme simplexovú hĺbku (obrázok 1.3), vidíme, že napriek tomu, že určité rozdiely medzi extrémnymi pozorovaniami z hľadiska oboch hĺbok sú, na pohľad sa oba výsledné obrázky vôbec nelíšia. Hlavný rozdiel je v počte rôznych úrovní hĺbky: zatiaľ čo polopriestorová hĺbka 51 pozorovaní roztriedila do 16 rôznych kategórií, simplexová hĺbka rovnaké pozorovania usporiadala až do 44 rôznych skupín. To naznačuje, že simplexová hĺbka je v istom zmysle citlivejšia ako polopriestorová.

Príklad 4. Pri použití simplexovej hĺbky na kontaminovaný náhodný výber z príkladu 2 dostávame opäť vizuálne takmer identické výsledky ako pri použití polopriestorovej hĺbky. Kontaminujúci bod je označený za odľahlý a dokonca aj pozorovania pochybne označené za extrémne podľa polopriestorovej hĺbky dostávajú nízku hodnotu simplexovej hĺbky. Výsledky teda dávajú tušiť, že plne nezapadajú do štruktúry náhodného výberu.

Poznamenajme ale na koniec, že v prípade iného ako normálneho rozdelenia, napríklad nejakého rozdelenia s nesymetrickým nosičom, by boli rozdiely medzi polopriestorovou a simplexovou hĺbkou oveľa výraznejšie. Hĺbky totiž ani pre malé dimenzie pozorovaných náhodných vektorov nemusia dávať rovnaké usporiadanie.



Obr. 1.3: Simplexová hĺbka a dvojrozmerné normálne rozdelenie.



Obr. 1.4: Simplexová hĺbka a kontaminované dvojrozmerné normálne rozdelenie.

Okrem dvoch spomenutých bolo zavedené množstvo ďalších hĺbok. Všeobecné vlastnosti, ktoré by štatistická hĺbková funkcia v konečnorozmernom priestore mala splňovať, vyšetrovali Zuo a Serfling [29]. Hĺbka by mala „zoradiť“ body podľa nejakej vzdialenosti od centra rozdelenia. Vo všeobecnom prípade nemusí byť jasné, čo má centrom rozdelenia byť, je to však zrejmé v prípade symetrických rozdelení. Definujme preto najprv symetrické rozdelenie pravdepodobnosti na vektorovom priestore.

Definícia:(Polopriestorovo symetrické rozdelenie)
Nech $P \in \mathcal{P}(M)$ a X je náhodná veličina s rozdelením pravdepodobnosti P. Potom rozdelenie tejto náhodnej veličiny je *polopriestorovo symetrické* okolo bodu $\theta \in M$, ak platí

$$P(X \in H) \geq \frac{1}{2}$$

pre každý uzavretý polopriestor $H \in \mathcal{H}(M)$ taký, že $\theta \in H$.

Súvislosť polopriestorovej symetrie s polopriestorovou hĺbkou 1.3 je zrejmá.

O niečo širšou definíciou symetrie rozdelenia pravdepodobnosti je angulárna symetria v tom zmysle, že každé polopriestorovo symetrické rozdelenie na normovanom vektorovom priestore je tiež angulárne symetrické, ale nie naopak. Oba druhy symetrie sú si však veľmi podobné a ako ukázal Serfling [25], oba koncepty symetrie sú v prípade absolútne spojitých rozdelení ekvivalentné. Zaveďme teda pre úplnosť ešte angulárne symetrické rozdelenia ako najširšiu triedu symetrických rozdelení.

Definícia:(Angulárne symetrické rozdelenie)
Nech X je náhodná veličina s rozdelením pravdepodobnosti na normovanom vektorovom priestore M nad $\mathbb{R}$ s normou $\|.\|$. Potom rozdelenie tejto náhodnej veličiny je *angulárne symetrické* okolo bodu $\theta \in M$, ak platí

$$\frac{(X-\theta)}{\|X-\theta\|} \stackrel{\mathcal{D}}{=} -\frac{(X-\theta)}{\|X-\theta\|},$$

kde $\stackrel{\mathcal{D}}{=}$ označuje rovnosť distribúcií náhodných veličín.

Zaveďme teraz štatistickú hĺbkovú funkciu pre $\mathbb{R}^d$ tak ako ju definovali Zuo a Serfling [29]. Pre náhodný vektor X budeme ako P_X označovať rozdelenie pravdepodobnosti vektoru X a pre merateľnú funkciu $T : \mathbb{R}^d \to \mathbb{R}^d$ ako $P_{T(X)}$ rozdelenie pravdepodobnosti transformovaného rozdelenia $T(X)$.

Definícia:(Štatistická hĺbková funkcia v $\mathbb{R}^d$)
Štatistická hĺbková funkcia v priestore $\mathbb{R}^d$ je také obmedzené a nezáporné zobrazenie

$$D(.;.) : \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1],$$

že

P1 rovnosť
$$D(Ax+b;P_{AX+b}) = D(x;P_X) \tag{1.5}$$
platí pre každé $P_X \in \mathcal{P}\left(\mathbb{R}^d\right)$, pre každé $x \in \mathbb{R}^d$, pre každú regulárnu maticu $A \in \mathbb{R}^{d\times d}$ a pre každý vektor $b \in \mathbb{R}^d$.

P2 $D(\theta;P) = \sup_{x\in\mathbb{R}^d} D(x;P)$ platí pre každé $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ také, že θ je stredom (angulárnej) symetrie rozdelenia P.

P3 pre každé $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ s bodom s najvyššou hodnotou hĺbky (najhlbším bodom) $\theta \in \mathbb{R}^d$ platí
$$D(x;P) \le D(\theta+\alpha(x-\theta);P)$$
pre každé $\alpha \in [0,1]$ a $x \in \mathbb{R}^d$.

P4 pre každé $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ platí $D(x;P) \to 0$ pre $\|x\| \to \infty$.

Podmienku P1 nazývame tiež podmienkou *afinnej invariancie*, P2 popisuje vlastnosť *maximality v strede symetrie*, P3 *monotóniu relatívnu voči bodu s najväčšou hĺbkou* a P4 podmienku *nulovosti limity*.

Pre neskoršie účely zaveďme slabšiu verziu podmienky afinnej invariancie 1.5. Funkcia $D(.;.) : \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]$ splňuje podmienku *slabej afinnej invariancie*, ak

P1b rovnosť

$$D(cx+b;P_{cX+b}) = D(x;P_X) \tag{1.6}$$

platí pre každé $P_X \in \mathcal{P}\left(\mathbb{R}^d\right)$, pre každé $x \in \mathbb{R}^d$, pre každé $c \in \mathbb{R}$, $c \neq 0$ a pre každý vektor $b \in \mathbb{R}^d$.

Samozrejme ak hĺbková funkcia splňuje P1, splňuje triviálne aj P1b.

Ako dokázali Zuo a Serfling [29], polopriestorová hĺbka je v konečnorozmernom prípade štatistickou hĺbkovou funkciou v zmysle poslednej definície. V článku je ďalej dokázané, že pre absolútne spojité angulárne symetrické rozdelenia na $\mathbb{R}^d$ je aj simplexová hĺbka štatistickou hĺbkovou funkciou. V prípade diskrétneho rozdelenia však už simplexová hĺbka nemusí splňovať podmienku maximality P2 ani monotónie P3, my sa však v ďalšom budeme často obmedzovať práve na absolútne spojité rozdelenia.

Uveďme teraz dve známe tvrdenia o vlastnostiach simplexovej hĺbky dát tak, ako ich neskôr využijeme pri ďalších dôkazoch.

Tvrdenie 1.1. *Nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ a X je náhodný vektor z rozdelenia P. Potom pre $SD(.;.)$ platí 1.5 pre každú regulárnu maticu $A \in \mathbb{R}^{d\times d}$ a pre každý vektor $b \in \mathbb{R}^d$.*

Dôkaz. Dôkaz je jednoduchý, vyplýva z toho, že afinnú transformáciu môžeme vyjadriť ako zloženie posunutia $T_b : x \mapsto x+b$ a lineárneho zobrazenia $T_A : x \mapsto Ax$. Invariancia voči posunutiu je zrejmá a invariancia voči lineárnemu zobrazeniu plynie z

$$T_A\left(\sum_{i=1}^{s} \xi_i x_i\right) = \sum_{i=1}^{s} \xi_i T_A(x_i) \tag{1.7}$$

pre každé $s \in \mathbb{N}$ a $\xi_i \in \mathbb{R}$, $i = 1,\ldots,s$. Preto ak bod $x \in \mathbb{R}^d$ leží v konvexnom obale (prípadne simplexe) bodov $x_1,\ldots,x_{d+1}$, teda

$$x = \sum_{i=1}^{d+1} \xi_i x_i$$

pre $\xi_i \in [0,1]$, podľa 1.7 platí

$$T_A(x) = \sum_{i=1}^{s} \xi_i T_A(x_i)$$

pre $\xi_i \in [0,1]$, a bod $T_A(x)$ teda leží v konvexnom obale bodov $T_A(x_i)$. Zložením dvoch invariantných zobrazení dostávame invariantné zobrazenie, čím je dôkaz ukončený. □

Tvrdenie 1.2. *Nech P je absolútne spojité rozdelenie pravdepodobnosti na $\mathbb{R}^d$ angulárne symetrické okolo bodu $\theta \in \mathbb{R}^d$. Potom pre každé $x \in \mathbb{R}^d$ je*

$$SD(x;P) \leq SD(\theta + \alpha(x-\theta);P)$$

pre každé $\alpha \in [0,1]$. Ďalej platí $SD(\theta;P) = 2^{-d}$.

Dôkaz. Pozri Liu [11]. □

V práci sa pokúsime modifikovať konečnorozmerné hĺbky tak, aby sa dali použiť na špeciálny typ dát, funkcionálne dáta.

1.2 Funkcionálne dáta

Pri opakujúcich sa meraniach jednej veličiny sa často stáva, že napozorujeme dáta, na ktoré je prirodzene výhodnejšie alebo vyslovene nutné nazerať ako na reálne funkcie pre spôsob, akým vznikajú. Ako príklad uveďme experiment, pri ktorom na $n \in \mathbb{N}$ pacientoch meriame v priebehu choroby opakovane telesnú teplotu. Inými príkladmi podobných dát môžu byť rôzne meteorologické merania ako merania veľkosti zrážok, sily vetra alebo teploty, hydrologické merania výšky hladín riek, merania priebehu rastu jedincov alebo celých populácií v čase, rovnako ako množstvo iných medicínskych alebo priemyselných meraní charakteristík meniacich sa v čase. Ramsay a Silverman [20, 21] popísali veľa podobných príkladov.

V našom experimente na i-tom pacientovi prevedieme $s_i \in \mathbb{N}$ meraní v časoch $t_{i1}, \ldots, t_{is_i}$, ktoré označíme ako $y_{i1}, \ldots, y_{is_i}$. Predpokladajme, že všetky časy, v ktorých boli dáta pozorované, sú body z nejakého uzavretého intervalu $I \subset \mathbb{R}$, ktorý môžeme interpretovať ako dobu trvania choroby u pacientov. V celej práci budeme bez ujmy na všeobecnosti za takýto interval brať $I = [0,1]$. Inak v prípade $I = [a,b]$, $a,b \in \mathbb{R}$ pristúpime k normujúcej transformácii času

$$T : t \longmapsto \frac{t-a}{b-a}.$$

Ak by sme sa na pozorovania prevedené na i-tom pacientovi pokúsili nazerať ako na s_i-rozmerný náhodný vektor $y_i = (y_{i1}, \ldots, y_{is_i})^T$, narazili by sme na závažný problém, pretože hodnota pozorovania y_{ij} pre $j = 1, \ldots, s_i$ závisí na čase t_{ij}, v ktorom bolo meranie prevedené. My sme však apriórne nepredpokladali, že by boli merania teploty prevedené u všetkých pacientov v rovnakých časoch, preto nemôžeme merania $y_{1j}, \ldots, y_{nj}$ považovať za náhodný výber, pretože pozorovania mohli byť prevedené v rôznych časoch. Dokonca sme ani nepredpokladali, že by pre každého pacienta bol prevedený rovnaký počet pozorovaní, a preto by sa vo všeobecnosti mohli dokonca líšiť dimenzie jednotlivých náhodných vektorov $y_1, \ldots, y_n$. Aj v ideálnom prípade $s_1 = s_2 = \ldots = s_n = s$ a $t_{1j} = t_{2j} = \ldots = t_{nj} = t_j$ pre každé $j = 1, \ldots, s$ sa však spomínaný prístup ukazuje ako nevhodné alebo aspoň ťažkopádne riešenie, pretože sada pozorovaní u jedného pacienta môže vykazovať určitý trend, ktorý je často dôležitejší ako samotné hodnoty vektoru pozorovaní. Pri snahe nahliadať na pozorovania ako na konečnorozmerné náhodné vektory by sme ale stratili možnosti inferencie trendu, ktorý pozorovania vykazujú.

Ako prirodzenejší a intuitívne jasnejší prístup sa javí nahliadať na sadu meraní u i-teho pacienta ako na s_i hodnôt nejakej nepozorovateľnej funkcie telesnej teploty pacienta $x_i : [0,1] \to \mathbb{R}$ v časoch $t_{i1}, \ldots, t_{is_i}$. Tieto funkčné hodnoty niekedy môžeme považovať za presné, teda nezaťažené chybou merania, čím pristupujeme na model

$$y_{ij} = x_i\left(t_{ij}\right), i = 1, \ldots, n, j = 1, \ldots, s_i.$$

V praxi však presné merania dostávame zriedka a preto budeme uvažovať prípad, kedy namerané pozorovanie y_{ij} považujeme za hodnotu funkcie x_i v čase t_{ij} zaťaženú nejakou náhodnou chybou ε_{ij}. Chyby budú tvoriť náhodný výber z rozdelenia s nulovou strednou hodnotou a spoločným konečným rozptylom. Tým prijímame model

$$y_{ij} = x_i\left(t_{ij}\right) + \varepsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, s_i. \tag{1.8}$$

Na takéto náhodné funkcie teploty jednotlivých pacientov $x_1, \ldots, x_n$ sa už môžeme pozerať ako na náhodný výber z rozdelenia pravdepodobnosti P na nejakom priestore M

funkcií s definičným oborom v intervale $[0,1]$. Takéto dáta, ktoré môžeme reprezentovať funkciou, nazývame *funkcionálne dáta*.

Aby sme tieto nepozorovateľné funkcie mohli odhadnúť na základe napozorovaných vektorov y_i, budeme od funkcionálneho priestoru M vyžadovať, aby splňoval niektoré prirodzené vlastnosti vyplývajúce z povahy experimentu.

Vo všetkých spomenutých príkladoch môžeme apriórne predpokladať, že funkcie tvoriace namerané pozorovania sú spojité, prípadne až do istej miery hladké. Ak zostaneme pri príklade s priebehom telesnej teploty u pacientov, tá sa zrejme tiež nebude meniť skokovite v priebehu času a rovnako ako väčšina procesov v prírode bude vykazovať aj istú hladkosť v priebehu. Preto sa v analýze funkcionálnych dát budeme obmedzovať na spojité funkcie, v niektorých prípadoch dokonca na hladké funkcie.

Iným argumentom v prospech toho, že je dôležité obmedzovať priestor funkcií M (napríklad na triedu spojitých funkcií) je to, že ak by sme nekládli na trajektórie náhodných procesov, akými tieto funkcie vlastne sú, žiadne predpoklady, dostali by sme niečo ako nekonečnorozmernú sadu náhodných veličín bez nejakého jednoduchého súvisu medzi nimi. Pretože ale môžeme vykonať v priebehu času reálne iba konečné množstvo pozorovaní funkčných hodnôt, nemali by sme žiadnu informáciu o priebehu funkcií v časoch medzi diskrétnymi časmi meraní a tak by sa celá analýza dát ako funkcií opäť musela redukovať na konečnorozmernú analýzu jednotlivých meraní.

Narazili sme na to, že v skutočnosti napriek tomu, že dáta považujeme za funkcie, nikdy ich nemôžeme pozorovať vo všetkých bodoch ich definičného oboru. Každá sada pozorovaní jedného objektu je teda podľa 1.8 iba aproximácia skutočnej funkčnej hodnoty nepozorovateľnej funkcie v niekoľkých diskrétnych časových okamihoch. Preto, aby sme sa na pozorovanie mohli skutočne pozerať ako na funkciu, musíme samotnú realizáciu náhodného procesu odhadnúť na základe pozorovaní v jednotlivých bodoch. Ako uvádzajú Ramsay a Silverman [21], toto je možné v podstate dvomi rôznymi spôsobmi:

- Vyhladzovanie bodov y_{ij} založené na parametrickej regresii (konečnorozmerný prípad)
- Lokálne vyhladzovanie alebo jadrové odhady x_i založené na neparametrickej regresii (nekonečnorozmerný prípad)

Podstatný rozdiel medzi týmito dvomi prístupmi je v odlišnosti priestoru M tvoreného funkciami, ktoré je možné takýmto vyhladením získať. Opustime teraz príklad s meraním telesnej teploty pacientov a zhrňme, akým spôsobom je možné postupovať pri aproximácii nejakej funkcie x.

Predpokladajme, že z povahy dát je zrejmé, že skutočná nepozorovateľná funkcia x, ktorú sa v prvom kroku snažíme odhadnúť, je diferencovateľná a má spojité derivácie až do rádu $K \in \mathbb{N}_0 \cup \{\infty\}$. Predpokladáme teda platnosť $M \subseteq C^{(K)}([0,1])$. V prvom prípade regresného vyhladenia aproximujeme funkciu x nejakou lineárnou kombináciou $d \in \mathbb{N}$ vhodne volených bázových funkcií $\{\varphi_i\}_{i=1}^{d} \subset C^{(K)}([0,1])$. Nájdeme tak priblíženie nameraných hodnôt funkciou z konečnorozmerného vektorového priestoru

$$\mathcal{L} \equiv \mathcal{L}(\varphi_1,\ldots,\varphi_d) = \left\{ \sum_{i=1}^{d} c_i \varphi_i(t) \in C^{(K)}([0,1]) : c = (c_1,\ldots,c_d)^T \in \mathbb{R}^d \right\}. \quad (1.9)$$

Jedná sa teda o funkcie jednoznačne dané bázovými funkciami a vektorom konštánt $c = (c_1,\ldots,c_d)^T \in \mathbb{R}^d$.

V druhom prípade lokálneho alebo jadrového vyhladzovania už dostávame nekonečnorozmerný, a teda ťažko popísateľný priestor funkcií. Z toho dôvodu už v ďalšom nebudeme môcť využívať jednoduché vlastnosti konečnorozmerných priestorov tak ako v prípade regresného vyhladzovania. Na takýto odhad funkcie x budeme nahliadať ako na funkciu $x \in C^{(K)}([0,1])$ bez možnosti nejakej reštrikcie na vlastný podpriestor.

Pri ďalšej analýze funkcionálnych dát sa stotožňuje nepozorovateľná funkcia x s jej odhadom založeným na vyhladzovaní.

Veľkou výhodou funkcionálneho poňatia dát je to, že pri predpoklade hladkosti je možné pracovať aj s tvarom samotnej funkcie a to tak, že je možné analýzu previesť nielen na funkčné hodnoty $x(t)$ v bodoch $t \in [0,1]$, ale aj na derivácie pozorovanej funkcie. Zaveďme konvenciu, že funkčnú hodnotu funkcie x v bode jej definičného oboru t budeme považovať za nultú deriváciu funkcie x v bode t, a označujme ďalej ako $x^{(k)}(t)$ k-tu deriváciu funkcie x v bode $t \in (0,1)$ pre $k = 0,\ldots,K$. Rozšírenie analýzy na všetky dostupné derivácie podstatne rozširuje možnosti inferencie funkcionálnych dát o možnosť skúmania tvaru funkcií. Ako však uvidíme, žiadna doposiaľ zavedená a používaná hĺbka pre funkcionálne dáta takýmto spôsobom tvar funkcie do výpočtu jej hĺbky nezahrňuje.

Uveďme však ešte funkcionálnu analógiu definície štatistickej hĺbkovej funkcie tak ako ju budeme neskôr používať. Nech $M \subset C([0,1])$ a $\|.\|$ je norma na M.

Definícia:(Štatistická hĺbková funkcia v $C([0,1])$)
Štatistická hĺbková funkcia v priestore $M \subset C([0,1])$ je také obmedzené a nezáporné zobrazenie

$$D(.;.) : M \times \mathcal{P}(M) \to [0,1],$$

že

P1 rovnosť

$$D(ax+b; P_{aX+b}) = D(x; P_X) \tag{1.10}$$

platí pre pre každé $P_X \in \mathcal{P}(M)$, pre každé $x \in M$ a pre všetky $a,b \in M$, kde $ax(t) = a(t)x(t)$ pre $t \in [0,1]$.

P2 $D(\theta; P) = \sup_{x \in M} D(x; P)$ platí pre každé $P \in \mathcal{P}(M)$ také, že θ je stredom (angulárnej) symetrie rozdelenia P.

P3 pre každé $P \in \mathcal{P}(M)$ s bodom s najvyššou hodnotou hĺbky (najhlbším bodom) $\theta \in M$ platí

$$D(x; P) \leq D(\theta + \alpha(x-\theta); P)$$

pre každé $\alpha \in [0,1]$ a $x \in M$.

P4 pre každé $P \in \mathcal{P}(M)$ platí $D(x; P) \to 0$ pre $\|x\| \to \infty$.

Kapitola 2

Funkcionálne hĺbky

2.1 Indukované hĺbky

V prípade, že nepozorovateľné funkcie stotožňujeme s funkciami získanými parametrickým regresným vyhladením diskrétnych pozorovaní, získavame vždy iba funkcionálne pozorovania z konečnorozmerného priestoru charakterizovaného množinou bázových funkcií $\{\varphi_i\}_{i=1}^d \subset C^{(K)}([0,1])$. Funkcie $\{\varphi_i\}_{i=1}^d$ budeme v celom ďalšom texte bez ujmy na všeobecnosti považovať za lineárne nezávislé. Jedná sa už teda o pozorovania z rozdelenia na vlastnom lineárnom podpriestore $\mathcal{L} \subset C^{(K)}([0,1])$ zavedenom v 1.9. Môžeme ich preto vzájomne jednoznačne stotožňovať s vektorom konštánt $c = (c_1, \ldots, c_d)^T \in \mathbb{R}^d$ vzhľadom k báze.

Predpokladajme ďalej, že aj funkcie, ktorých hĺbku budeme voči rozdeleniu na priestore $\mathcal{L}$ merať, budú vyhladené rovnakým spôsobom ako všetky ostatné funkcie, a teda budú prvkami $\mathcal{L}$.

Týmto sa však ale môžeme v celom ďalšom postupe obmedziť na takýto konečnorozmerný podpriestor a ak použijeme vetu o izomorfizme konečnedimenzionálnych vektorových priestorov, môžeme stotožniť priestor $\mathcal{L}$ s priestorom $\mathbb{R}^d$ izomorfizmom daným vzťahom

$$\psi : x(t) = \sum_{i=1}^{d} c_i \varphi_i(t) \longmapsto (c_1, \ldots, c_d)^T . \tag{2.1}$$

Pretože zobrazenie ψ zachováva všetky „podstatné“ vlastnosti oboch priestorov, je možné prejsť izomorfne do priestoru $\mathbb{R}^d$, vyčísliť hĺbku funkcie $\sum_{i=1}^d c_i \varphi_i(t)$ na jej koeficientoch vzhľadom k báze $(c_1, \ldots, c_d)^T$ voči rozdeleniu $P_\psi = \psi(P)$ na $\mathbb{R}^d$ a funkcii priradiť hĺbku jej koeficientov voči rozdeleniu indukovanému zobrazením ψ.

Na základe takejto úvahy môžeme definovať jednu triedu hĺbok pre funkcionálne dáta z konečnorozmerných podpriestorov priestorov funkcií.

Definícia:(Indukovaná funkcionálna hĺbka)
Nech $D : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to [0,1]$ je štatistická hĺbková funkcia a nech P je rozdelenie pravdepodobnosti na d-rozmernom vektorovom priestore funkcií $\mathcal{L} \subset C([0,1])$ danom bázou $\{\varphi_i\}_{i=1}^d \subset C([0,1])$. Potom *hĺbka indukovaná štatistickou hĺbkovou funkciou D voči rozdeleniu pravdepodobnosti P* je zobrazenie

$$D_{\mathcal{L}}(.;P) : \mathcal{L} \to [0,1]$$

také, že pre $x(t) = \sum_{i=1}^d c_i \varphi_i(t)$ je

$$D_{\mathcal{L}}(x;P) = D\left(\psi(x); P_\psi\right). \tag{2.2}$$

Vyšetrime teraz, za akých podmienok bude $D_{\mathcal{L}}$ skutočne štatistickou hĺbkovou funkciou.

Tvrdenie 2.1. *Nech P je rozdelenie pravdepodobnosti na unitárnom priestore $\mathcal{L} \subset C([0,1])$ generovanom bázou funkcií $\{\varphi_i\}_{i=1}^d$. Nech zobrazenie $D : \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]$ je štatistická hĺbková funkcia. Potom zobrazenie $D_{\mathcal{L}}(.;P) : \mathcal{L} \to [0,1]$ definované v 2.2 je štatistickou hĺbkovou funkciou na priestore $\mathcal{L}$.*

Dôkaz. Nech $D(.;P)$ je štatistická hĺbková funkcia a $x(t) = \sum_{i=1}^d c_i\varphi_i(t) \in \mathcal{L}$. Stotožnime priestory $\mathcal{L}$ a $\mathbb{R}^d$ zobrazením $\psi : \mathcal{L} \to \mathbb{R}^d$ zavedeným v 2.1. Afinné transformácie $\mathcal{L}$ budú teda priamo afinné transformácie koeficientov funkcií $c \in \mathbb{R}^d$. Preto afinnou transformáciou funkcie x pomocou zobrazenia $T_{A,b}$ určeného regulárnou maticou $A \in \mathbb{R}^{d\times d}$ a vektorom $b \in \mathbb{R}^d$ je funkcia $g \in \mathcal{L}$ taká, že

$$g(t) = \sum_{i=1}^d v_i\varphi_i(t) + \sum_{i=1}^d b_i\varphi_i(t),$$

kde $v = (v_1,\ldots,v_d)^T = Ac$. Potom platí

$$D_{\mathcal{L}}\left(T_{A,b}(x);T_{A,b}(P)\right) = D\left(Ac+b;\psi\left(T_{A,b}(P)\right)\right) = D\left(c;P_\psi\right) = D_{\mathcal{L}}(x;P) \qquad (2.3)$$

pre každú regulárnu maticu $A \in \mathbb{R}^{d\times d}$ a pre každý vektor $b = (b_1,\ldots,b_d)^T \in \mathbb{R}^d$. Predpoklad afinnej invariancie P1 je teda pre $D_{\mathcal{L}}(.;P)$ triviálne splnený.

V ďalšom budeme bez ujmy na všeobecnosti predpokladať, že báza priestoru $\mathcal{L}$ je ortonormálna voči skalárnemu súčinu na $\mathcal{L}$. To je možné, pretože podľa 2.3 je indukovaná hĺbka funkcií invariantná voči afinnej, a teda aj lineárnej transformácii vektorov koeficientov voči báze. Lineárna transformácia vektoru koeficientov c voči bázovým funkciám pomocou regulárnej matice A však nie je nič iné ako vyjadrenie vektoru c voči transformovanej báze funkcií priestoru $\mathcal{L}$ danej maticou prechodu A. Grammova-Schmidtova ortonormalizácia bázy funkcií $\{\varphi_i\}_{i=1}^d$ je tiež lineárna transformácia, a preto bude hĺbka funkcie x voči ľubovoľnej báze $\{\varphi_i\}_{i=1}^d$ rovnaká ako indukovaná hĺbka x voči ortonormalizovanej báze priestoru $\mathcal{L}$.

Vyšetrujme teraz podmienku P2. Za stred symetrie budeme považovať stred angulárnej symetrie a skalárny súčin na priestore $\mathcal{L}$ označíme $\langle .,. \rangle_{\mathcal{L}}$. Ukážeme, že za uvedených predpokladov sú skalárny súčin na vektoroch koeficientov funkcií a skalárny súčin $\langle .,. \rangle_{\mathcal{L}}$ totožné. Pre funkciu $g(t) = \sum_{i=1}^d v_i\varphi_i(t)$ platí

$$\langle x,g\rangle_{\mathcal{L}} = \left\langle \sum_{i=1}^d c_i\varphi_i, \sum_{j=1}^d v_j\varphi_j \right\rangle_{\mathcal{L}} = \sum_{i=1}^d\sum_{j=1}^d c_iv_j\left\langle \varphi_i,\varphi_j\right\rangle_{\mathcal{L}} = c^Tv. \qquad (2.4)$$

Zobrazenie ψ je teda unitárny izomorfizmus priestorov $\mathcal{L}$ a $\mathbb{R}^d$, a preto stredy symetrie v oboch priestoroch budú pomocou zobrazenia ψ sebe zodpovedajúce body. Podmienka P2 je teda splnená.

Pretože zobrazenie ψ je izomorfizmus a teda zachováva lineárnu štruktúru, podmienka monotónie relatívnej voči najhlbšiemu bodu P3 v definícii štatistickej hĺbkovej funkcie je tiež triviálne splnená.

Nakoniec podmienka limitného správania hĺbkovej funkcie P4 je rovnako splnená, pretože podľa 2.4 platí $\|f\|_{\mathcal{L}} = \|c\|$ a norma v oboch priestoroch je totožná.

Preto zobrazenie $D_{\mathcal{L}}(.;P)$ indukované štatistickou hĺbkovou funkciou D pomocou zobrazenia ψ je štatistickou hĺbkovou funkciou v priestore $\mathcal{L}$. □

Konštrukciou 2.2 teda dostávame pre každú hĺbku v $\mathbb{R}^d$ nejakú hĺbku v d-rozmernom priestore funkcií.

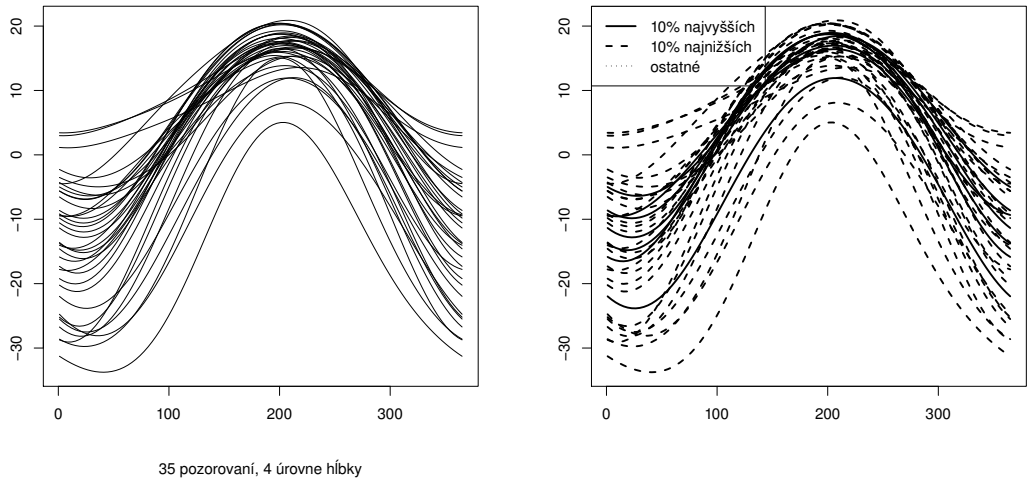
Napriek tomu, že podľa tvrdenia 2.1 je indukovaná hĺbka konečnorozmerných funkcionálnych dát skutočne štatistickou hĺbkovou funkciou, použitie takejto triedy hĺbkových funkcionálov je veľmi limitované. To je dané najmä tým, že väčšinou býva dimenzia priestoru, do ktorého sa funkcie regresne vyhladzujú relatívne vysoká voči počtu pozorovaných funkcií v náhodnom výbere. Preto typickým javom, ktorý pri počítaní takejto hĺbky nastáva, a zároveň najväčší problém v praktickom použití indukovaných hĺbok je to, že veľká časť, prípadne dokonca všetky funkcie náhodného výberu dostanú nulovú hĺbku. To je spôsobené tým, že v priestore vysokej dimenzie je jednoducho príliš málo pozorovaní na rozlíšenie typických od odľahlých. Tento jav si ilustrujeme na dvoch príkladoch reálnych a simulovaných dát.

Najprv však uveďme niekoľko technických detailov. Všetky výpočty vo všetkých príkladoch boli vykonané v programe `R 2.11.1` ([19]). Hĺbka dvojrozmerných dát a hĺbka viacrozmerných koeficientov funkcionálnych dát v prípade indukovaných hĺbok bola počítaná pomocou procedúr z balíka `depth` ([18]). V ostatných príkladoch funkcionálnych dát boli funkcie prevedené do diskretizovanej podoby v 101 (v kapitole 5 do 51) bodoch mriežky (ekvidistantných bodoch definičného oboru). To znamená, že sme každú funkciu interne reprezentovali ako 101-rozmerný vektor funkčných hodnôt na mriežke. V kapitole 4 budeme využívať hodnoty prvej derivácie funkcií. Tie vždy získavame pomocou vhodnej metódy numerického diferencovania (závisiacej od samotnej metódy vyhladzovania funkcií) a následne opäť diskretizujeme do 101 bodov mriežky. Hĺbky funkcií boli pre urýchlenie výpočtov počítané pomocou procedúr naprogramovaných v jazyku `C++` (pozri prílohu A pre zdrojové kódy) a následne použitých volaním do programu `R` (pozri prílohu B).

Príklad 5. Majme náhodný výber funkcionálnych dát pozostávajúci z priemerných teplôt v jednotlivých dňoch roka v 35 kanadských mestách (pozri Ramsay a Silverman [20]). Pre každé mesto regresne vyhladíme týchto 365 diskrétnych pozorovaní do priestoru dimenzie 5 trigonometrických polynómov rádu 2. Hĺbku takýchto dát spočítame pomocou indukovanej polopriestorovej hĺbky. Na obrázku 2.1 máme v ľavej časti znázornené tieto vyhladené pozorovania. V pravej časti je zvýraznených 10 % pozorovaní s najvyššou a najnižšou hodnotou hĺbky voči náhodnému výberu podobne ako v predchádzajúcich príkladoch.

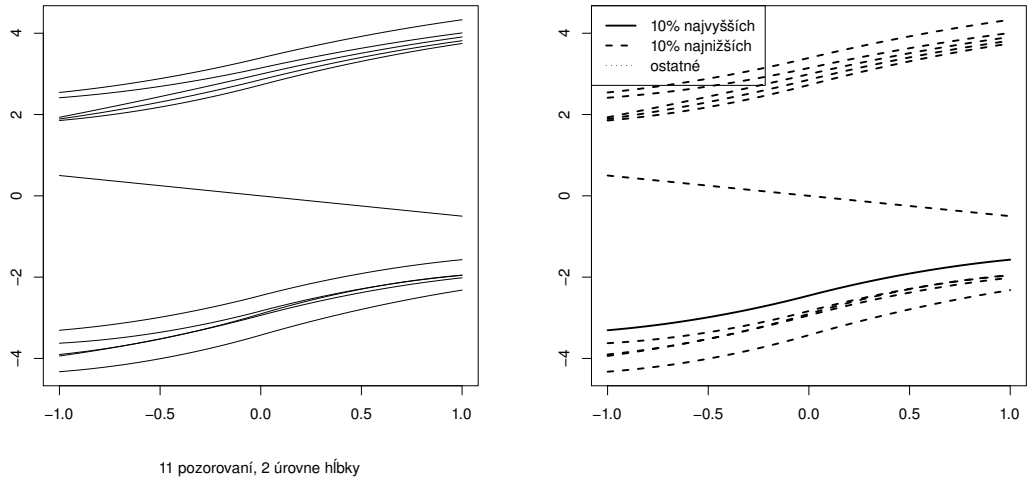
Pozorovania boli podľa rastúcej hĺbky rozdelené do štyroch skupín, veľká väčšina ich však dostala najmenšiu, nulovú hĺbku. Napriek tomu, že skupina funkcií s najväčšou hĺbkou zodpovedá našej predstave o funkciách s „typickým priebehom“, všetky funkcie s najnižšou hĺbkou určite nemôžeme prehlásiť za kandidátov na odľahlé pozorovania.

Príklad 6. Majme náhodný výber simulovaných dát, ktorý tvoria dve skupiny rýdze rastúcich funkcií a jedna netypická klesajúca funkcia. Dáta boli vyhladené do priestoru kubických B-spline funkcií s jedným uzlovým bodom uprostred intervalu $[-1, 1]$. Na obrázku 2.2 sú rovnako ako v príklade 5 znázornené najprv tieto vyhladené pozorovania a následne zvýraznené pozorovania s extrémnymi hodnotami indukovanej polopriestorovej hĺbky vzhľadom k náhodnému výberu. Funkcií je však príliš málo na to, aby vôbec boli rozlíšené pomocou indukovanej hĺbky, a tak boli pozorovania roztriedené iba do dvoch skupín, pričom jediná funkcia dostala nenulovú hodnotu hĺbky.



Obr. 2.1: Indukovaná polopriestorová hĺbka a kanadské počasie.

Poznamenajme však, že táto funkcia s najväčšou hodnotou hĺbky sa skutočne javí ako typická v náhodnom výbere.



Obr. 2.2: Indukovaná polopriestorová hĺbka a simulované dáta.

Ako sme videli na príkladoch, napriek tomu, že indukovaná hĺbka dát je štatistická hĺbková funkcia, v mnohých prípadoch je prakticky nepoužiteľná. Jednoznačnou nevýhodou je obmedzenie sa na konečnorozmerné funkcionálne priestory, no ani v tomto prípade nemusí dávať indukovaná hĺbka dobré výsledky. S rastúcou dimenzionalitou pozorovaní je totiž príliš citlivá na presnosť funkcionálnych pozorovaní, čo vo veľkej časti prípadov nie je možné zaručiť vzhľadom na to, že sme v prvom kroku stotožnili nepozorovateľné funkcie s vyhladenými diskrétnymi pozorovaniami.

Kapitola 3

Geometrické hĺbky

Zaoberajme sa teraz iným prístupom k určovaniu hĺbky funkcionálnych dát. Najprv definujeme a niektoré základné vlastnosti uvedieme pre najjednoduchšiu, *Fraimanovu-Munizovej hĺbku* pre funkcionálne dáta. Napriek tomu, že jej použitie je veľmi obmedzené a často dáva neuspokojivé výsledky, neskôr niektoré jej vlastnosti použijeme pri konštrukcii zovšeobecnenia iných hĺbok. Ďalej zavedieme tzv. *pásové hĺbky* pre funkcionálne dáta, ktoré sú založené na grafickej reprezentácii funkcií a určujú hĺbku funkcie podľa toho, ako veľmi jej graf leží medzi grafmi ostatných funkcií náhodného výberu, alebo ako veľmi sa na grafy ostatných funkcií náhodného výberu podobá. Skôr ako prejdeme k funkcionálnym dátam, zavedieme pásovú hĺbku v jednoduchšom, konečnorozmernom prípade, pretože ako neskôr uvidíme, pásové hĺbky funkcionálnych dát sú iba priamym zovšeobecnením pásových hĺbok konečnorozmerných dát, pričom niektoré z vlastností konečnorozmerných pásových hĺbok sa priamo prenášajú aj na funkcionálny prípad.

3.1 Fraimanova-Munizovej hĺbka

Fraimanov a Munizovej článok [10] je možné považovať za prvý systematický výskum možností rozšírenia hĺbky dát na funkcionálny prípad. Je v ňom odvodený jednoduchý hĺbkový funkcionál založený na tom, že danú jednorozmernú hĺbkovú funkciu tvaru 1.1 je možné upraviť tak, aby merala hĺbku voči rozdeleniu s nosičom v súčinovom priestore S s ľubovoľnou dimenzionalitou. Postupuje sa tak, že na priestor S nazeráme ako na kartézsky súčin jednorozmerných podpriestorov

$$S = \prod_{\lambda \in \Lambda} S_\lambda,$$

pričom marginálne rozdelenie P na priestore S_λ označíme ako P_λ a projekciu bodu $s \in S$ na podpriestor S_λ označíme ako s_λ. Potom pre bod $s \in S$ vyhodnotíme čiastočné hĺbky $D_\lambda = D(s_\lambda; P_\lambda)$ pre $\lambda \in \Lambda$ a nakoniec ich jednoducho vyintegrujeme cez množinu Λ (prípadne sčítame alebo spriemerujeme pre spočetný alebo konečný prípad). Tým dostávame *Fraimanovu-Munizovej hĺbku indukovanú hĺbkovou funkciou D*. Takúto hĺbku samozrejme môžeme použiť v priestore spojitých funkcií $C([0,1])$.

Definícia:(Fraimanova-Munizovej hĺbka pre funkcionálne dáta)
Nech $P \in \mathcal{P}(C([0,1]))$ a $x \in C([0,1])$. Označme pre $t \in [0,1]$ ako $P_t^{(0)}$ jednorozmerné marginálne rozdelenie funkčnej hodnoty funkcií z rozdelenia P v bode t. Potom

Fraimanova-Munizovej hĺbka funkcie x indukovaná hĺbkou D vzhľadom k rozdeleniu pravdepodobnosti P je definovaná ako

$$DF^D(x;P) = \int_0^1 D\left(x(t);P_t^{(0)}\right)\mathrm{d}t. \tag{3.1}$$

Výberová verzia Fraimanovej-Munizovej hĺbky je zavedená ako Fraimanova-Munizovej hĺbka pre empirické rozdelenie na priestore $C([0,1])$ a tak bude záležať na výberovej verzii jednorozmernej hĺbky D. Fraiman a Muniz [10] skúmali podrobnejšie Fraimanovu-Munizovej hĺbku indukovanú simplexovou a polopriestorovou hĺbkou. Za istých technických predpokladov na nosič $E \subset C([0,1])$ rozdelenia P bolo pre Fraimanovu-Munizovej hĺbku indukovanú simplexovou a polopriestorovou hĺbkou dokázané dôležité tvrdenie o silnej rovnomernej konzistencii na množine E, pod ktorou budeme rozumieť vlastnosť

$$\sup_{x\in E}|D_n(x) - D(x)| \xrightarrow[n\to\infty]{\text{s.i.}} 0, \tag{3.2}$$

kde D_n označuje výberovú verziu hĺbky D.

V článku bola ďalej zavedená jednoduchá aplikácia hĺbkovej funkcie pre funkcionálne dáta, najjednoduchší L-odhad strednej hodnoty, α-useknutý priemer funkcionálnych dát. Pre náhodný výber $X_1,\ldots,X_n \in E$ (v našom prípade uvažujeme $E \subset C([0,1])$) sa jedná o aritmetický priemer $n - \lfloor n\alpha \rfloor$ pozorovaní s najvyššou hodnotou hĺbky v náhodnom výbere, kde $\lfloor a \rfloor$ je najväčšie $m \in \mathbb{Z}$ také, že $m \le a$. Jedná sa teda o aritmetický priemer takých pozorovaní, ktorých hĺbka je vyššia ako nejaká daná konštanta β. Definujme teraz takýto robustný odhad parametru polohy ako ho uviedli Fraiman a Muniz [10].

Definícia:(β-useknutý priemer-populačná verzia)
Nech $P \in \mathcal{P}(E)$, $D : E \times \mathcal{P}(E) \to [0,1]$ je hĺbková funkcia a

$$\beta \in \left[0, \sup_{x\in E} D(x;P)\right].$$

Nech $\mathrm{E}[\mathbb{I}[D(X;P) \ge \beta]] > 0$. Potom β-*useknutý priemer* rozdelenia pravdepodobnosti P definujeme ako

$$\bar{X}_\beta = \frac{\mathrm{E}[\mathbb{I}[D(X;P) \ge \beta]X]}{\mathrm{E}[\mathbb{I}[D(X;P) \ge \beta]]}. \tag{3.3}$$

Podobne môžeme definovať výberovú verziu β-useknutého priemeru.

Definícia:(β-useknutý priemer-výberová verzia)
Nech $X_1,\ldots,X_n$ je náhodný výber z $P \in \mathcal{P}(E)$, $D_n : E \times E^n \to [0,1]$ je výberová verzia hĺbkovej funkcie a

$$\beta \in \left[0, \sup_{x\in E} D_n(x;X_1,\ldots,X_n)\right].$$

Nech $\sum_{i=1}^n \mathbb{I}[D_n(X_i;X_1,\ldots,X_n) \ge \beta] > 0$.Potom β-*useknutý priemer* funkcií takéhoto náhodného výberu definujeme ako

$$\widehat{\overline{X}_{n\beta}} = \frac{\sum_{i=1}^n \mathbb{I}[D_n(X_i;X_1,\ldots,X_n) \ge \beta]X_i}{\sum_{i=1}^n \mathbb{I}[D_n(X_i;X_1,\ldots,X_n) \ge \beta]}. \tag{3.4}$$

V našom prípade je tvrdenie zaujímavé najmä pre prípad funkcionálneho výberového priestoru $E \subseteq C([0,1])$, resp. neskôr aj $E \subseteq C^{(K)}([0,1])$ pre nejaké $K \in \mathbb{N}$.

V prípade $\beta = 0$ sa v tomto prípade jedná o neuseknutý aritmetický priemer funkcionálnych pozorovaní, v druhom krajnom prípade $\beta = \sup_{x \in E \subseteq C([0,1])} D(x;P)$ sa jedná o zovšeobecnený medián funkcionálnych dát.

Za predpokladu silnej rovnomernej konzistencie hĺbkovej funkcie je možné dokázať silnú konzistenciu β-useknutého priemeru.

Tvrdenie 3.1. *Nech $P \in \mathcal{P}(E)$ (nie nutne $E \subset C([0,1])$). Nech D, resp. D_n sú populačná a výberová verzia štatistiky na E také, že platí 3.2. Potom výberová verzia β-useknutého priemeru 3.4 je silne konzistentným odhadom populačnej verzie 3.3, to znamená*

$$\widehat{\overline{X}_{n\beta}} \xrightarrow[n\to\infty]{s.i.} \overline{X}_{\beta}.$$

Dôkaz. Pozri Fraiman a Muniz [10]. □

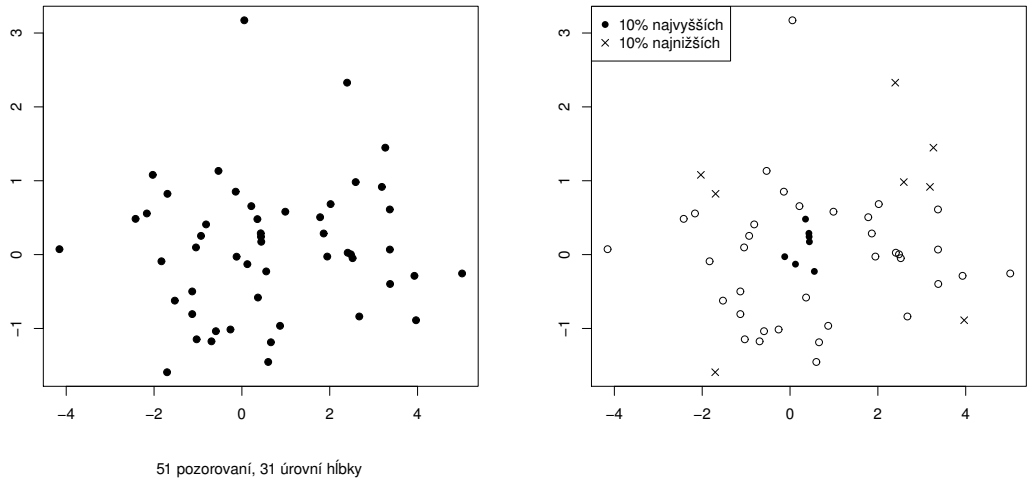
Napriek tomu, že Fraimanova-Munizovej hĺbka je za istých predpokladov rovnomerne silne konzistentná, rozhodne sa nejedná o hĺbkový funkcionál s dobrými vlastnosťami, čo naznačuje už jej verzia pre prípad konečnorozmerného rozdelenia. Ak by sme Fraimanovu-Munizovej hĺbku zavádzali na priestore konečnej dimenzie d, bude sa pre bod $x \in \mathbb{R}^d$ jednať iba o priemer (prípadne súčet) jednorozmerných hĺbok daných projekciou do jednotlivých jednorozmerných marginálnych rozdelení. Skúsme teda Fraimanovu-Munizovej hĺbku použiť na známe príklady z kapitoly 1.

Príklad 7. Uvažujme dva náhodné výbery z dvojrozmerného normálneho rozdelenia z príkladov 1 a 2, resp. 3 a 4. Počítajme dvojrozmernú Fraimanovu-Munizovej hĺbku indukovanú (napríklad) polopriestorovou hĺbkou každého bodu oboch náhodných výberov voči zvyšným pozorovaniam.

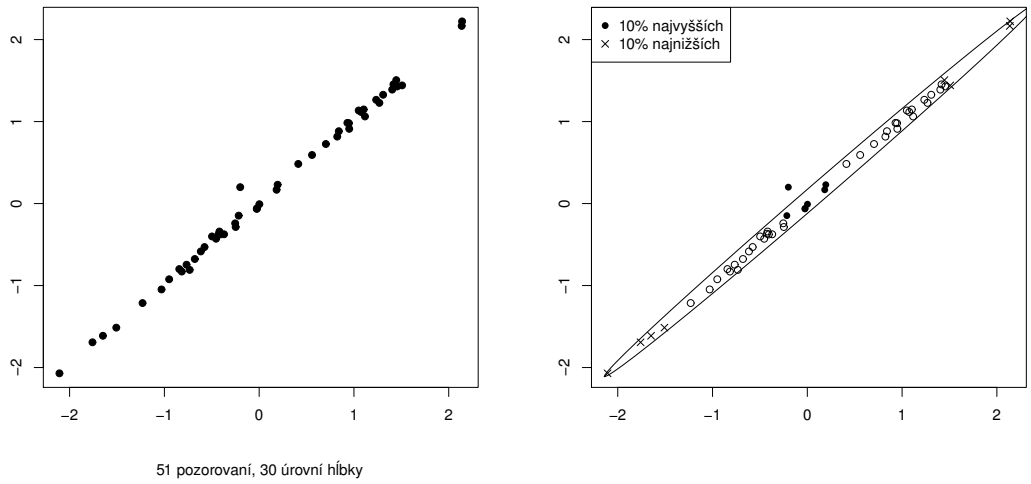
V prípade nekorelovaného náhodného výberu (obrázok 3.1) vidíme, že napriek tomu, že rozdiely oproti hĺbkam z kapitoly 1 sú viditeľné, nezdajú sa byť nijako významné. Hlavné rozdiely sú v extrémnych pozorovaniach v niektorej zo súradníc. Preto napríklad pozorovanie s výrazne najnižšou hodnotou X-ovej zložky bolo zhodne podľa polopriestorovej aj simplexovej hĺbky identifikované ako kandidát na odľahlý bod, zatiaľ čo pomerne typická hodnota Y-ovej zložky mu zaručuje vysokú hodnotu Fraimanovej-Munizovej hĺbky.

Tento jav je ešte lepšie vidieť na príklade s korelovaným normálnym rozdelením (obrázok 3.2). Kontaminujúci bod, napriek tomu, že leží mimo 99% konfidenčnú elipsu rozdelenia, nie je identifikovaný ako odľahlé pozorovanie, a dokonca nadobúda veľmi vysokú hodnotu hĺbky. To je spôsobené práve tým, že v rámci oboch zložiek posudzovaných osobitne sú jeho súradnice pomerne typické voči náhodnému výberu a tým aj súčet čiastočných hĺbok súradníc tohto bodu bude vysoký.

Najväčšou slabosťou Fraimanovej-Munizovej hĺbky a vôbec všetkých hĺbkových funkcií založených na projekciách do priestorov nižšej dimenzie je neschopnosť odlíšenia bodov štrukturálne sa vymykajúcich náhodnému výberu. Inými slovami, Fraimanova-Munizovej hĺbka nemá vlastnosť afinnej invariancie P1, ale iba vlastnosť slabej afinnej invariancie P1b. Práve afinná invariancia zaručuje, že bod vymykajúci sa (eliptickej) štruktúre náhodného výberu je správne identifikovaný ako kandidát na odľahlé pozorovanie. Vlastnosti afinnej invariancie Fraimanovej-Munizovej hĺbky vo všeobecnejšom prípade neskôr vyplynú z tvrdenia 3.4.



Obr. 3.1: Fraimanova-Munizovej hĺbka a dvojrozmerné normálne rozdelenie.



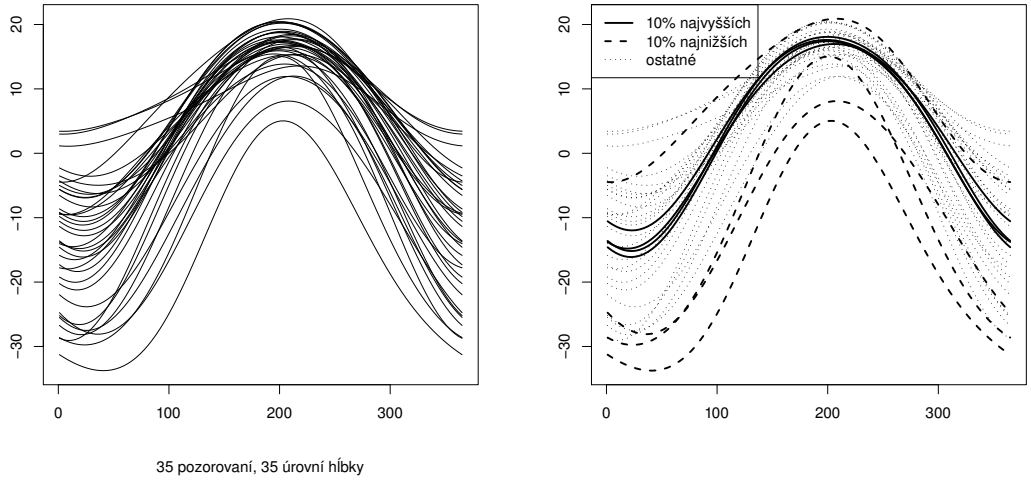
Obr. 3.2: Fraimanova-Munizovej hĺbka a kontaminované dvojrozmerné normálne rozdelenie.

Záverom príkladu teda zhrňme dôležité pozorovanie, že aj odľahlé body náhodného výberu môžu byť pri analýze pomocou Fraimanovej-Munizovej hĺbky mylne považované za veľmi typické.

Napriek tomu, že v niektorých prípadoch dáva už pre konečnorozmerné rozdelenia Fraimanova-Munizovej hĺbka zlé výsledky, ukážme si jej použitie ešte na niekoľkých príkladoch funkcionálnych dát.

Príklad 8. Pri použití Fraimanovej-Munizovej simplexovej hĺbky na kanadské teplotné dáta popísané v príklade 5 dostávame na pohľad veľmi dobré výsledky, ako vidíme na obrázku 3.3. Takmer rovnaké výsledky by sme dostali použitím Fraimanovej-Munizovej polopriestorovej hĺbky. Typické funkcie s grafmi výlučne uprostred zhluku pozorovaní dostávajú najväčšie hodnoty hĺbky, zreteľne odľahlé funkcie najmenšie hodnoty hĺbky. Rozlíšenie funkcií z hľadiska počtu úrovní hĺbky je výborné: dosiahli

sme dokonalé rozlíšenie 35 funkcií, pri použití polopriestorovej hĺbky dosahujeme podobný výsledok 33 z 35. Je teda jasné, že aj tak jednoduchý hĺbkový funkcionál ako Fraimanova-Munizovej hĺbka môže pri jednoduchých sadách pozorovaní dávať (zdanlivo) dobré výsledky.

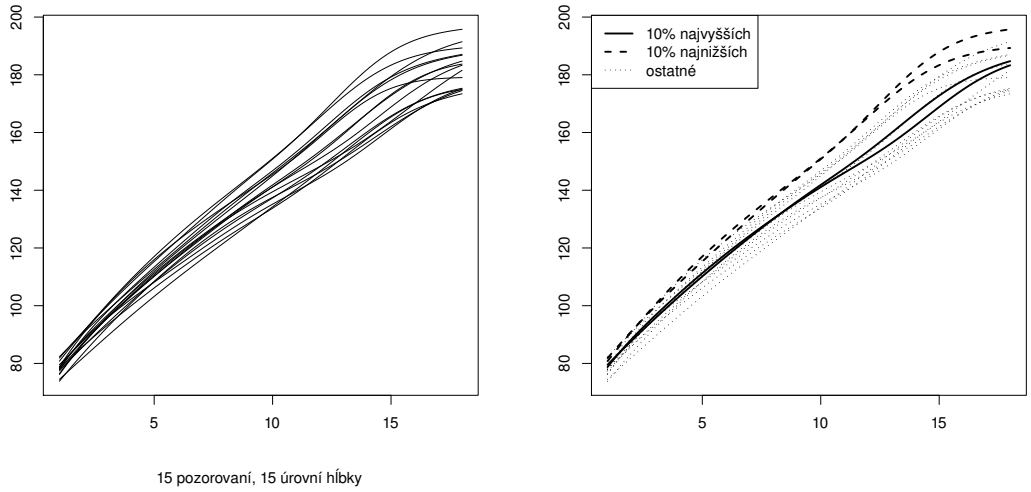


Obr. 3.3: Fraimanova-Munizovej hĺbka a kanadské počasie.

Príklad 9. Použime ďalej Fraimanovu-Munizovej simplexovú hĺbku na náhodný výber rastových kriviek. Jedná sa o klasické dáta pochádzajúce zo štúdie rastu detí Univerzity v Berkeley (pozri Tuddenham a Snyder [27] alebo podobná štúdia od Falknera [7], podrobne popísané Ramsaym a Silvermanom [20, 21]). Pozorovania sú funkcie výšky 54 dievčat a 39 chlapcov, každé merané 29-krát od narodenia do osemnástych narodenín. Pre jednoduchosť budeme v našom texte vždy pracovať iba s 15 pozorovaniami náhodne vybranými z rastových kriviek chlapcov. Dáta sme previedli do funkcionálnej podoby metódou vyhladzovania kubickými lokálnymi polynómami a nejedná sa preto o náhodný výber dát z konečnorozmerného priestoru. Na ich analýzu sme teda nemohli použiť indukovanú hĺbku.

Pri použití Fraimanovej-Munizovej simplexovej hĺbky (obrázok 3.4) vidíme rovnako ako v príklade 8 dokonalé rozlíšenie funkcií a na pohľad dobré rozlíšenie typických a odľahlých pozorovaní. Veľmi podobné výsledky by opäť dávala ako Fraimanova-Munizovej polopriestorová, tak aj každá iná Fraimanova-Munizovej hĺbka indukovaná hĺbkovou funkciou s dobrými vlastnosťami.

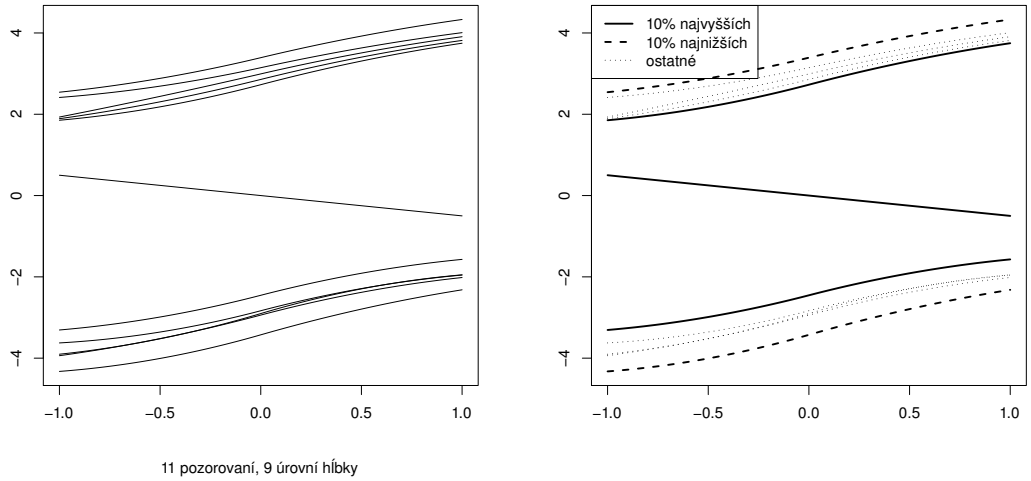
Príklad 10. Aby sme ukázali funkcionálnu analógiu chyby z príkladu 7, ktorej sa Fraimanova-Munizovej hĺbka bude dopúšťať pri netypickom rozdelení dát, použime Fraimanovu-Munizovej simplexovú hĺbku na náhodný výber simulovaných dát z príkladu 6 (obrázok 3.5). Budeme sa snažiť úspešne identifikovať jedinú klesajúcu funkciu v náhodnom výbere rastúcich tak ako sa to podarilo indukovaným hĺbkam (príklad 6). Jedná sa totiž podobne ako v príklade 7 o jasne netypické pozorovanie voči náhodnému výberu napriek tomu, že leží uprostred dvoch zhlukov funkcií. V prípade Fraimanovej-Munizovej hĺbky však práve klesajúca funkcia, aj keď sa tvarom odlišuje od ostatných, dostáva najvyššiu hodnotu hĺbky a preto by sa malo jednať o ekvivalent



Obr. 3.4: Fraimanova-Munizovej hĺbka a rast chlapcov.

zovšeobecneného mediánu pre náš náhodný výber. To však zrejme neplatí, pretože ako výberový medián by sme nemali označovať funkciu iba preto že leží medzi dvomi zhlukmi, ale preto, že má v istom zmysle podobné vlastnosti ako funkcie náhodného výberu. Mediánová funkcia by mala byť typická funkcia najlepšie popisujúca rozdelenie pravdepodobnosti. To však ale jediná klesajúca funkcia v náhodnom výbere rastúcich byť nemôže.

Ako sme videli, Fraimanove-Munizovej hĺbky zlyhávajú, ak majú odlíšiť v náhodnom výbere funkcionálnych dát pozorovania odľahlé v tvare funkcií také, že zároveň nie sú odľahlé v polohe.



Obr. 3.5: Fraimanova-Munizovej hĺbka a simulované dáta.

Fraimanove-Munizovej hĺbky dosahovali (zdanlivo) dobré výsledky v príkladoch 8, 9, ale napríklad aj v článku Febrera et al. [8], kde pomocou Fraimanovej-Munizovej hĺbky autori dobre identifikovali odľahlé pozorovania v náhodnom výbere funkcií kon-

centrácie oxidov dusíka v atmosfére v priebehu času. Dôvodom toho je, že v jednoduchom, nepatologickom prípade s veľkým rozsahom výberu sú pozorovania odľahlé v tvare väčšinou tiež pozorovaniami odľahlými v polohe, pretože v náhodnom výbere je jednoducho príliš veľa podobných funkcií blízko seba a tak pozorovanie netypické v tvare koliduje s ostatnými. V horších prípadoch s malým počtom pozorovaní sa však môže stať, že Fraimanove-Munizovej hĺbky úplne zlyhajú a ako najtypickejšieho reprezentanta náhodného výberu vyberú funkciu odľahlú v tvare. Tento problém sa netýka iba jednoduchých Fraimanových-Munizovej hĺbok. Ako ukážeme v časti 3.3, rovnaký problém s odlíšením odľahlých pozorovaní určitého typu budú mať aj iné v literatúre odporúčané hĺbkové funkcionály.

Vlastnosti Fraimanovej-Munizovej hĺbky nebudeme v tejto chvíli podrobnejšie vyšetrovať, pretože ako neskôr uvidíme, jedná sa iba o špeciálny prípad obsiahnutý v širšej triede zovšeobecnených pásových funkcionálov, ktoré detailnejšie popíšeme v časti 3.3.

3.2 Pásové hĺbky pre konečnorozmerné dáta

López-Pintado a Romo [15, 16] zaviedli a skúmali *pásové hĺbky* pre body z d-rozmerného euklidovského priestoru $\mathbb{R}^d$ a následne pre funkcionálne dáta z priestoru spojitých funkcií $C([0,1])$. Pásová hĺbka má v konečnorozmernom prípade oproti iným používaným hĺbkovým funkciám veľkú výhodu toho, že je jednoducho implementovateľná a výpočetne rýchla aj pre vysoké hodnoty dimenzie d (pozri kapitolu 6). Je založená na koncepte *pásu* tvoreného bodmi z $\mathbb{R}^d$.

Definícia:(Pás v $\mathbb{R}^d$)
Pre body $x_1,\ldots,x_n \in \mathbb{R}^d$ definujme *pás tvorený bodmi* $x_1,\ldots,x_n$ ako

$$R(x_1,\ldots,x_n) = \left\{ x \in \mathbb{R}^d : \min_{i=1,\ldots,n} x_i(k) \leq x(k) \leq \max_{i=1,\ldots,n} x_i(k), \forall k = 1,\ldots,d \right\},$$

kde $x(k)$ označuje hodnotu k-tej zložky vektoru $x \in \mathbb{R}^d$. Jedná sa teda o d-rozmerný interval medzi minimálnymi a maximálnymi súradnicami bodov v každej zložke.

Teraz môžeme zaviesť pásovú hĺbku pre konečnorozmerné dáta tak ako López-Pintado a Romo [15, 16].

Definícia:(Pásová hĺbka v $\mathbb{R}^d$-populačná verzia)
Nech $x \in \mathbb{R}^d$ a $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ a $X_1,\ldots,X_J$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$LP^j(x;P) = P\left(x \in R\left(X_1,\ldots,X_j\right)\right). \tag{3.5}$$

Potom definujeme *d-dimenzionálnu pásovú hĺbku J-teho rádu bodu x vzhľadom k rozdeleniu pravdepodobnosti P* ako

$$LP^{(J)}(x;P) = \frac{1}{J-1} \sum_{j=2}^{J} LP^j(x;P).$$

V diskrétnom prípade náhodného výberu z konečnorozmerného rozdelenia pravdepodobnosti zavádzame výberovú verziu pásovej hĺbky ako príslušnú U-štatistiku.

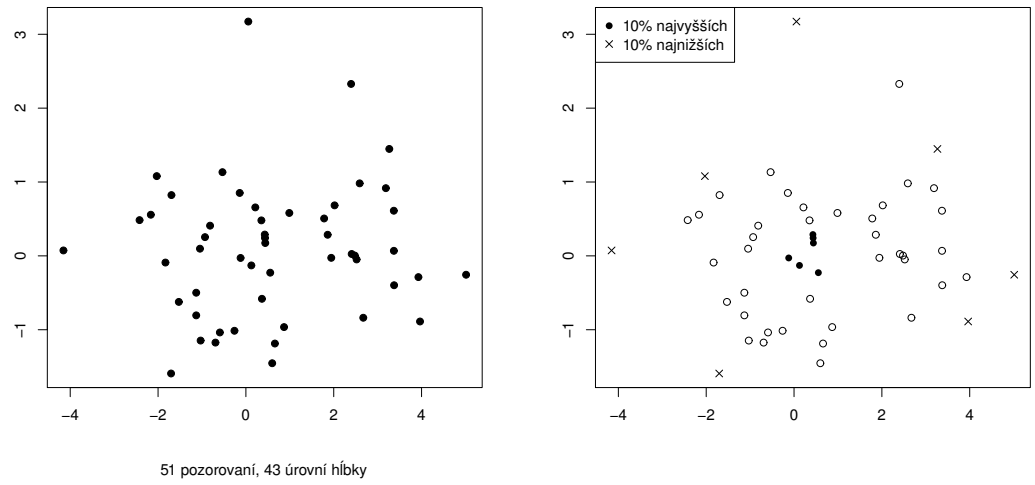
Definícia:(Pásová hĺbka v $\mathbb{R}^d$-výberová verzia)
Nech $x \in \mathbb{R}^d$ a $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$, $n \geq J$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$LP_n^j(x;\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \mathbb{I}\left[x \in R\left(X_{i_1},\ldots,X_{i_j}\right)\right]. \tag{3.6}$$

Potom definujeme *d-dimenzionálnu pásovú hĺbku J-teho rádu bodu x vzhľadom k náhodnému výberu* $\mathbb{X}$ ako

$$LP_n^{(J)}(x;\mathbb{X}) = \frac{1}{J-1}\sum_{j=2}^{J} LP_n^j(x;\mathbb{X}).$$

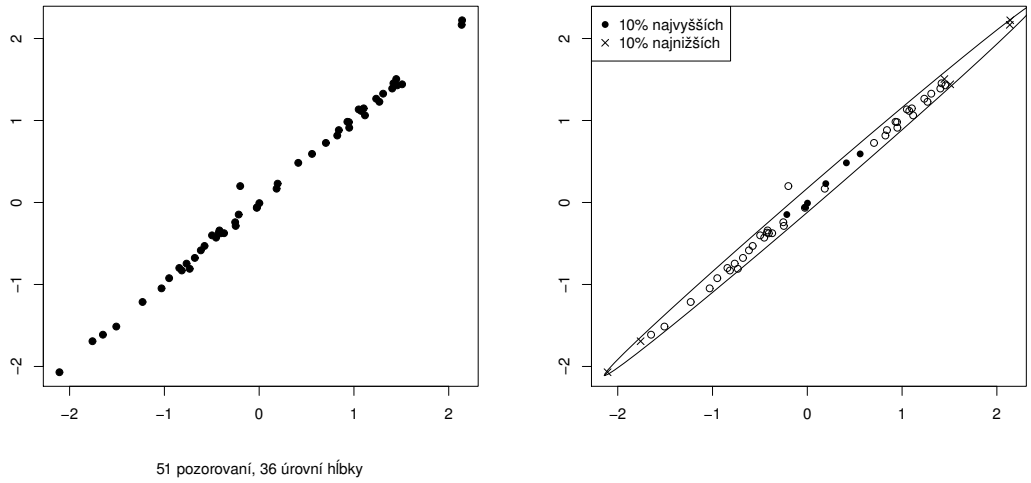
Príklad 11. Rovnako ako v príkladoch 1, 2, 3, 4 a 7 skúsme aplikovať na oba náhodné výbery z dvojrozmerného normálneho rozdelenia pásovú hĺbku druhého rádu. Na obrázkoch 3.6 a 3.7 vidíme, že podobne ako simplexová hĺbka aj pásová hĺbka dobre rozlišuje body, pretože usporiadala náhodný výber až do 43 (resp. 36) skupín rôznych úrovní hĺbky. Samozrejme ako pre všetky hĺbkové funkcie, najhlbšie sú body blízko centra náhodného výberu a najmenej hlboké body na okrajoch náhodného výberu. Rozdiely oproti simplexovej aj polopriestorovej hĺbke sú minimálne.



Obr. 3.6: Pásová hĺbka rádu 2 a dvojrozmerné normálne rozdelenie.

Najzaujímavejšie je, že kontaminujúci bod korelovaného rozdelenia nie je identifikovaný ani ako vyslovene typický ani ako odľahlý. Pri bližšom skúmaní môžeme zistiť, že jeho hodnota hĺbky je vyššia ako hodnota hĺbky viac ako 74 % ostatných pozorovaní. Pásová hĺbka je teda v tomto zmysle medzistupňom medzi afinne invariantnými a Fraimanovými-Munizovej hĺbkami. Ako však dokážeme v tvrdení 3.4, ani pásová hĺbka nie je afinne invariantná.

Obdobne ako pásovú hĺbku môžeme definovať *zovšeobecnenú pásovú hĺbku* v $\mathbb{R}^d$. Jediným rozdielom je, že namiesto indikátoru toho, či bod $x \in \mathbb{R}^d$, ktorého hĺbku hľadáme, leží vo všetkých súradniciach v páse tvorenom niektorými bodmi náhodného výberu, uvažujeme pomer počtu takých zložiek bodu, ktoré v tomto páse ležia voči celkovému počtu zložiek bodu x. Tento pomer môžeme interpretovať aj ako mieru takých zložiek x, ktoré ležia v páse tvorenom bodmi náhodného výberu.



Obr. 3.7: Pásová hĺbka rádu 2 a kontaminované dvojrozmerné normálne rozdelenie.

Definícia:(Zovšeobecnená pásová hĺbka v $\mathbb{R}^d$-populačná verzia)
Nech $x \in \mathbb{R}^d$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ a $X_1,\ldots,X_J$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$GLP^j(x;P) = \frac{1}{d}\sum_{k=1}^{d} P\left(\min_{r=1,\ldots,j} X_r(k) \le x(k) \le \max_{r=1\ldots,j} X_r(k)\right). \tag{3.7}$$

Potom *zovšeobecnená d-dimenzionálna pásová hĺbka J-teho rádu bodu x vzhľadom k rozdeleniu pravdepodobnosti P* je

$$GLP^{(J)}(x;P) = \frac{1}{J-1}\sum_{j=2}^{J} GLP^j(x;P).$$

V prípade náhodného výberu postupujeme analogicky ako v prípade pásovej hĺbky.

Definícia:(Zovšeobecnená pásová hĺbka v $\mathbb{R}^d$-výberová verzia)
Nech $x \in \mathbb{R}^d$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$, $n \ge J$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$GLP_n^j(x;\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\le i_1<\ldots<i_j\le n} \frac{1}{d}\sum_{k=1}^{d} \mathbb{I}\left[\min_{r=i_1,\ldots,i_j} X_r(k) \le x(k) \le \max_{r=i_1\ldots,i_j} X_r(k)\right]. \tag{3.8}$$

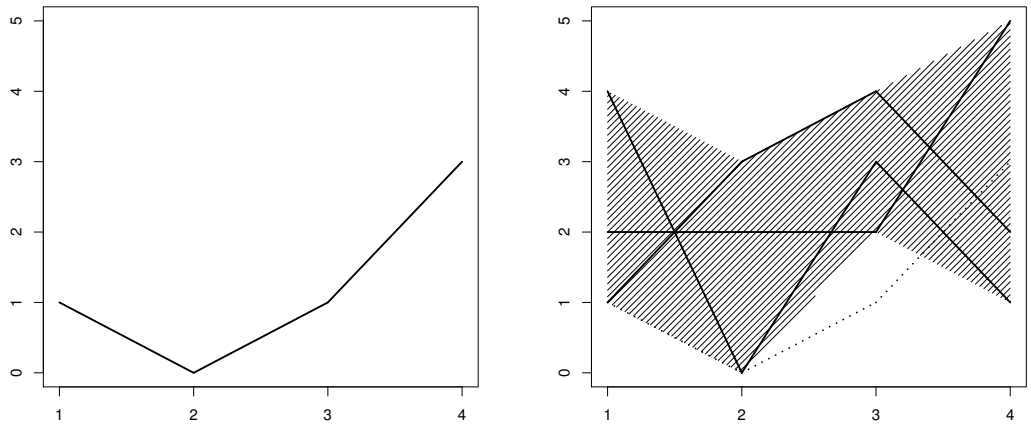
Potom *zovšeobecnená d-dimenzionálna pásová hĺbka J-teho rádu bodu x vzhľadom k náhodnému výberu* $\mathbb{X}$ je

$$GLP_n^{(J)}(x;\mathbb{X}) = \frac{1}{J-1}\sum_{j=2}^{J} GLP_n^j(x;\mathbb{X}).$$

V prípade všetkých pásových hĺbok sa ako parameter rádu J väčšinou v praxi z dôvodu vysokej výpočetnej náročnosti (pozri kapitolu 6) volí $J = 2$, 3 alebo v krajnom prípade 4, my však pre teoretické účely budeme uvažovať všetky možné voľby J.

Pretože zovšeobecnená pásová hĺbka rádu $J = 2$ aj $J = 3$ je ekvivalentná Fraimanovej-Munizovej simplexovej hĺbke (viac kapitola 6), v príkladoch použitia zovšeobecnenej pásovej hĺbky sa môžeme odkázať na príklad 7.

Aby sme lepšie videli prečo pásové hĺbky prirodzene využívajú geometrické vlastnosti súradnicového systému v $\mathbb{R}^d$, budeme pre zobrazovanie konečnorozmerných dát vyšších dimenzií používať *paralelný súradnicový systém*. V princípe sa jedná o to, že zložky d-rozmerného reálneho vektoru znázorníme na d rovnobežných rovnako vzdialených osiach. Pre $i = 1,\ldots,d$ vynášame na i-tu os hodnotu i-tej súradnice vektoru. Napríklad, bod $(1,0,1,3)^T \in \mathbb{R}^4$ môžeme v tomto prípade znázorniť ako na prvej časti obrázku 3.8.



Obr. 3.8: Bod $(1,0,1,3)^T \in \mathbb{R}^4$ a pás znázornené v paralelných súradniciach.

V takomto súradnicovom systéme je tiež pohodlné vykresľovať pásy v $\mathbb{R}^d$, pretože minimálna i maximálna zložka utvárajúca pás v každej súradnici je jednoducho znázorniteľná. Rovnako je možné ľahko z takéhoto grafu zistiť mieru náležania bodu do pásu. Ako príklad si uveďme znázornenie pásu tvoreného bodmi

$$x_1 = (1,3,4,2)^T,$$
$$x_2 = (2,2,2,5)^T,$$
$$x_3 = (4,0,3,1)^T.$$

Graficky vyhodnotíme mieru náležania bodu $x = (1,0,1,3)^T$ do takéhoto pásu (druhá časť obrázku 3.8, pás je znázornený šrafovaním). Z obrázku je jasne vidieť, že bod x leží v páse $B(x_1,x_2,x_3)$ v prvej, druhej a štvrtej zložke, v tretej zložke naopak v páse neleží, platilo by teda

$$\mathbb{I}[x \in R(x_1,\ldots,x_4)] = 0,$$
$$\frac{1}{4}\sum_{k=1}^{4} \mathbb{I}\left[\min_{r=1,\ldots,3} x_r(k) \leq x(k) \leq \max_{r=1,\ldots,3} x_r(k)\right] = \frac{3}{4}.$$

Znázornenie konečnorozmerných bodov v paralelných súradniciach je teda praktické a prirodzené najmä pre účely počítania d-dimenzionálnej pásovej hĺbky.

Zároveň sa však ukazujú možnosti zovšeobecnenia konečnorozmerných pásových hĺbok na funkcionálny prípad, pretože prirodzeným zovšeobecnením paralelných súradníc na funkcionálne alebo aj iné nekonečnorozmerné dáta je pripustenie nekonečne mnohých paralelných osí. K takémuto zovšeobecneniu môžeme pristúpiť v oboch nekonečnorozmerných prípadoch spočetnej aj nespočetnej mohutnosti počtu dimenzií. Grafické znázornenie funkcií v paralelných súradniciach je prirodzené, pretože sa jedná o bežné zobrazenie reálnej funkcie $x : [0,1] \to \mathbb{R}$ pomocou jej grafu v kartézskej sústave súradníc v $\mathbb{R}^2$.

Rozšírením konečnorozmernej pásovej hĺbky na reálne funkcie s kompaktným nosičom, teda nespočetnorozmerné dáta, dostaneme práve funkcionálnu pásovú hĺbku ako ju popíšeme v časti 3.3.

Konečnorozmernú zovšeobecnenú pásovú hĺbku nie je možné rozšíriť na funkcionálne dáta z dôvodu nekonečnej mohutnosti nosiča. Môžeme však zaviesť nejakú pravdepodobnostnú mieru $P_{[0,1]}$ na množine $[0,1]$ a následne zovšeobecnenú pásovú hĺbku funkcionálnych dát definovať ako $P_{[0,1]}$-mieru takých bodov, ktoré ležia v páse tvorenom niektorými funkciami náhodného výberu. Obe verzie pásovej hĺbky pre funkcionálne dáta takýmto spôsobom definujeme v časti 3.3.

Ukážme teraz niektoré vlastnosti pásových hĺbok v $\mathbb{R}^d$. Triviálne pozorovanie je, že všetky konečnorozmerné pásové hĺbky sú U-štatistiky.

Tvrdenie 3.2. *Nech $J = 2,3,\ldots$, $n \geq J$ a $x \in \mathbb{R}^d$. Potom $LP_n^{(J)}(x;.)$ aj $GLP_n^{(J)}(x;.)$ sú U-štatistiky rádu J.*

Dôkaz. López-Pintado a Romo [16] dokázali, že ak $H^2,\ldots,H^J$ sú U-štatistiky rádov $\{2,\ldots,J\}$, potom

$$H_J(x_1,\ldots,x_n) = \sum_{j=2}^{J} H^j(x_1,\ldots,x_n) \tag{3.9}$$

je U-štatistika rádu J. Zrejme však v našom prípade sú obe pásové hĺbky až na multiplikatívnu konštantu súčtom takýchto U-štatistík, v prípade $LP_n^{(J)}$ sa jedná o sčítance

$$LP_n^j(x;x_1,\ldots,x_n) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \mathbb{I}\left[x \in R\left(x_{i_1},\ldots,x_{i_j}\right)\right],$$

pre pevné j U-štatistiky rádu j s jadrom

$$h_j\left(x_1,\ldots,x_j\right) = \mathbb{I}\left[x \in R\left(x_1,\ldots,x_j\right)\right],$$

a podobne pre $GLP_n^{(J)}$ sa jedná o sčítance

$$GLP_n^j(x;x_1,\ldots,x_n) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \frac{1}{d}\sum_{k=1}^{d} \mathbb{I}\left[\min_{r=i_1,\ldots,i_j} x_r(k) \leq x(k) \leq \max_{r=i_1\ldots,i_j} x_r(k)\right]$$

rovnako U-štatistiky rádu j s jadrom

$$h_j^G\left(x_1,\ldots,x_j\right) = \frac{1}{d}\sum_{k=1}^{d} \mathbb{I}\left[\min_{r=1,\ldots,j} x_r(k) \leq x(k) \leq \max_{r=1\ldots,j} x_r(k)\right].$$

Preto je týmto podľa 3.9 tvrdenie dokázané. □

Venujme sa teraz vyšetrovaniu ďalších prirodzených vlastností, ktoré by rozumná hĺbková funkcia mala splňovať, medzi inými aj podmienku spojitosti. Pre tento účel si pripomeňme definíciu zhora a zdola polospojitých funkcií.

Definícia:(Polospojitá funkcia)
Nech M je topologický priestor a $x : M \to \mathbb{R}$ je nejaká funkcia na priestore M. Hovoríme, že x je *zhora (resp. zdola) polospojitá v bode* $t_0 \in M$, ak platí podmienka

$$\limsup_{t \to t_0} x(t) \leq x(t_0), \tag{3.10}$$

resp.

$$\liminf_{t \to t_0} x(t) \geq x(t_0). \tag{3.11}$$

Funkcia x je *zhora (zdola) polospojitá na na množine* $Y \subset M$, ak je zhora (zdola) polospojitá v každom bode $t \in Y$.

Pre pásovú hĺbku v $\mathbb{R}^d$ platí nasledujúca veta, ktorá zhŕňa nakoľko je takto definovaná funkcia skutočne štatistickou hĺbkovou funkciou v zmysle definície v časti 1.1.

Tvrdenie 3.3. *Nech P je absolútne spojité rozdelenie pravdepodobnosti na $\mathbb{R}^d$.*

- *Nech marginálne rozdelenia P sú symetrické okolo bodu $\theta \in \mathbb{R}^d$ a navyše hustota rozdelenia P je kladná na nejakom okolí θ. Potom $LP^{(J)}(x;P)$ je maximalizovaná jediným bodom, a to práve bodom θ.*

- *Nech marginálne rozdelenia P sú symetrické okolo bodu $\theta \in \mathbb{R}^d$. Potom*

$$LP^{(J)}(x;P) \leq LP^{(J)}(\theta + \alpha(x-\theta);P)$$

  pre každé $\alpha \in [0,1]$ a $x \in \mathbb{R}^d$.

- *Ak $\mathbb{X}_n = (X_1, \ldots, X_n)^T$ je náhodný výber z rozdelenia P, potom platí*

$$\sup_{\|x\|_\infty \geq M} LP^{(J)}(x;P) \xrightarrow[M\to\infty]{} 0,$$
$$\sup_{\|x\|_\infty \geq M} LP_n^{(J)}(x;\mathbb{X}_n) \xrightarrow[M\to\infty]{s.i.} 0$$

- *Potom $LP^{(J)}(x;P)$ je zhora polospojitá funkcia na $\mathbb{R}^d$. Ak navyše marginálne rozdelenia P sú absolútne spojité, potom $LP^{(J)}(x;P)$ je spojitá funkcia na $\mathbb{R}^d$.*

Dôkaz. Pozri López-Pintado a Romo [16]. □

Podľa tvrdenia 3.3 teda za technických predpokladov absolútnej spojitosti rozdelenia, symetrie marginálnych rozdelení a nenulovosti hustoty na nejakom okolí bodu stredu symetrie platia podmienky P2, P3 a P4 z definície štatistickej hĺbkovej funkcie. Ukážeme však, že podmienka afinnej invariancie P1 pre pásovú hĺbku vo viacrozmerných priestoroch neplatí.

Tvrdenie 3.4. *Nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ a $x \in \mathbb{R}^d$. Potom platí:*

- *pásová hĺbka je slabo afinne invariantná, to znamená platí 1.6.*
- *pre $d = 1$ je pásová hĺbka afinne invariantná, to znamená platí 1.5.*
- *pre $d > 1$ nie je pásová hĺbka afinne invariantná okrem degenerovaného prípadu rozdelenia pravdepodobnosti s nosičom v jednorozmernom podpriestore generovanom jedným z vektorov priestoru $\mathbb{R}^d$.*

Dôkaz. Prvá časť tvrdenia o slabej afinnej invariancii plynie z toho, že minimálne a maximálne súradnice množiny bodov $R \subset \mathbb{R}^d$ zostávajú ekvivariantné či už pri násobení reálnym číslom $c \neq 0$, ako aj pri pričítaní vektoru $b \in \mathbb{R}^d$. Pre minimálnu a maximálnu súradnicu k-tej zložky vektorov platí

$$\begin{aligned}
c \min_{X \in R} X(k) + b(k) &= \min_{X \in cR+b} X(k) && , c \geq 0, \\
c \max_{X \in R} X(k) + b(k) &= \max_{X \in cR+b} X(k) && , c \geq 0, \\
c \min_{X \in R} X(k) + b(k) &= \max_{X \in cR+b} X(k) && , c \leq 0, \\
c \max_{X \in R} X(k) + b(k) &= \min_{X \in cR+b} X(k) && , c \leq 0.
\end{aligned}$$

Dvojprvková množina minimálnej a maximálnej súradnice tak zostáva pri slabej afinnej transformácii ekvivariantná. Z toho plynie, že pásy zostávajú ekvivariantné, a preto aj všetky pásové hĺbky zostávajú slabo invariantné.

Druhá časť tvrdenia pre jednorozmerný prípad triviálne plynie z prvej, pretože v jednorozmernom prípade sú afinné transformácie totožné so slabými afinnými transformáciami, pre ktoré sme platnosť invariancie práve dokázali.

Dokážme teda nakoniec, že pre $d > 1$ neplatí afinná invariancia. Obmedzme sa na rozdelenia pravdepodobnosti, ktoré spĺňajú podmienku nedegenerovanosti nosiča uvedenú v tvrdení. Ak totiž nosičom rozdelenia bude jednorozmerný podpriestor daný niektorým bázovým vektorom priestoru $\mathbb{R}^d$, jedná sa zrejme opäť o prípad úplne analogický prípadu $d = 1$ a preto je v tomto prípade pásová hĺbka opäť afinne invariantná.

Afinná transformácia je zložením posunutia o vektor $b \in \mathbb{R}^d$ a lineárnej transformácie danej regulárnou maticou $A \in \mathbb{R}^{d \times d}$. Pásová hĺbka je však voči posunutiu o vektor b invariantná podľa prvej časti tvrdenia pre voľbu $c = 1$ v 1.6. Dokážme teda, že pásová hĺbka nie je lineárne invariantná, to znamená že existuje regulárna matica $A \in \mathbb{R}^{d \times d}$ a vektor $x \in \mathbb{R}^d$ tak, že

$$LP^{(J)}(Ax; P_{AX}) \neq LP^{(J)}(x; P_X). \tag{3.12}$$

To je ekvivalentné tvrdeniu, že existuje $j = 2, \ldots, J$, pás tvorený j bodmi v $\mathbb{R}^d$ a bod $x \in \mathbb{R}^d$, pre ktorý neplatí ekvivalencia

$$x \in R\left(X_1, \ldots, X_j\right) \Longleftrightarrow Ax \in R\left(AX_1, \ldots, AX_j\right) \tag{3.13}$$

pre nejakú A regulárnu. Majme nejakú množinu bodov $X_1, \ldots, X_j$ takú, že

$$R\left(X_1, \ldots, X_j\right) \supsetneq \mathrm{Conv}\left(X_1, \ldots, X_j\right), \tag{3.14}$$

kde $\mathrm{Conv}\left(X_1, \ldots, X_j\right)$ označuje uzavretý konvexný obal bodov $X_1, \ldots, X_j$. Za predpokladu nedegenerovanosti nosiča rozdelenia bude určite aspoň pre voľbu $j = 2$ s nenulovou pravdepodobnosťou existovať náhodný výber X_1, X_2 z rozdelenia P_X s vlastnosťou 3.14. Pre bod

$$x \in R\left(X_1, \ldots, X_j\right) \setminus \mathrm{Conv}\left(X_1, \ldots, X_j\right)$$

teraz nájdeme regulárnu maticu A tak, že neplatí 3.13.

K tomu stačí ukázať, že existuje taká regulárna matica A, pre ktorú

$$Ax \notin R\left(AX_1,\ldots,AX_j\right). \tag{3.15}$$

Pretože $x \notin \text{Conv}\left(X_1,\ldots,X_j\right)$, podľa Hahn-Banachovej vety o oddeliteľnosti bodu a kompaktnej množiny existuje lineárna forma $L:\mathbb{R}^d \to \mathbb{R}$ taká, že

$$Lx < \min_{y\in \text{Conv}\left(X_1,\ldots,X_j\right)} Ly \leq \min_{i=1,\ldots,j} LX_i. \tag{3.16}$$

Zostrojme teraz maticu lineárneho zobrazenia A tak, že d-rozmerný vektor L reprezentujúci lineárnu formu L dosadíme za prvý riadok matice A a ostatné riadky dodefinujeme ako ľubovoľných $(d-1)$ vektorov dopĺňajúcich bázu priestoru $\mathbb{R}^d$. Potom podľa 3.16 máme pre prvú zložku lineárne transformovaných bodov

$$Ax(1) = Lx < \min_{i=1,\ldots,j} LX_i = \min_{i=1,\ldots,j} AX_i(1),$$

z čoho plynie platnosť 3.15, a teda aj to, že neplatí 3.13. Tým sme dokázali, že v prípade, že sa niektorý pás tvorený náhodným výberom z rozdelenia P nezhoduje s konvexným obalom tohto náhodného výberu, existuje bod, pre ktorý sa nezachováva afinná invariancia pásovej hĺbky. □

Niekoľko ďalších zaujímavých tvrdení o pásových hĺbkach v $\mathbb{R}^d$ dokázali López-Pintado a Romo [16].

Dôkaz tvrdenia 3.4 nám tiež ukazuje možnosť, ako opraviť pásové hĺbky tak, aby boli afinne invariantné. Stačí totiž namiesto „slabo afinného“ pásu používať „afinné“ konvexné obaly. To nás vedie k nasledujúcej sérii definícií *konvexných hĺbok* ako afinne invariantných analógií pásových hĺbok.

Definícia:(Konvexná hĺbka v $\mathbb{R}^d$-populačná verzia)
Nech $x \in \mathbb{R}^d$ a $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ a $X_1,\ldots,X_J$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$CD^j(x;P) = P\left(x \in \text{Conv}\left(X_1,\ldots,X_j\right)\right). \tag{3.17}$$

Potom definujeme potom *d-dimenzionálnu konvexnú hĺbku J-teho rádu bodu x vzhľadom k rozdeleniu pravdepodobnosti P* ako

$$CD^{(J)}(x;P) = \frac{1}{J-1}\sum_{j=2}^{J} CD^j(x;P). \tag{3.18}$$

V prípade náhodného výberu z rozdelenia pravdepodobnosti P môžeme rovnako ako v prípade pásovej hĺbky zadefinovať výberovú verziu pomocou príslušnej U-štatistiky.

Definícia:(Konvexná hĺbka v $\mathbb{R}^d$-výberová verzia)
Nech $x \in \mathbb{R}^d$ a $J = 2,3,\ldots$, nech $P \in \mathcal{P}\left(\mathbb{R}^d\right)$, $n \geq J$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$CD_n^j(x;\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \mathbb{I}\left[x \in \text{Conv}\left(X_{i_1},\ldots,X_{i_j}\right)\right]. \tag{3.19}$$

Potom definujeme potom *d-dimenzionálnu konvexnú hĺbku J-teho rádu bodu x vzhľadom k náhodnému výberu* $\mathbb{X}$ ako

$$CD_n^{(J)}(x;\mathbb{X}) = \frac{1}{J-1}\sum_{j=2}^{J} CD_n^j(x;\mathbb{X}). \tag{3.20}$$

Konvexná hĺbka však nie je nič nové, pretože v špeciálnom prípade $J = d+1$ pre absolútne spojité rozdelenia je ekvivalentná simplexovej hĺbke definovanej v časti 1.1. V absolútne spojitom prípade sú totiž všetky členy CD^j a CD_n^j pre $j = 2,\ldots,d$ rovné nule. To platí, pretože pravdepodobnosť javu, že bod $x \in \mathbb{R}^d$ padne do konvexného obalu náhodného výberu j bodov z absolútne spojitého rozdelenia, ktorý má d-rozmernú Lebesgueovu mieru 0, je zrejme nulová. Pre prípad absolútne spojitých rozdelení je teda pri konvexnej hĺbke J-teho rádu zbytočné zahŕňať do výpočtu sčítance CD^j a CD_n^j pre $j = 2,\ldots,d$. Preto by sme mohli zjednodušiť 3.18 ako

$$CD^{(J)}(x) = \frac{1}{J-1}\sum_{j=d+1}^{J} CD^j(x),$$

respektíve 3.20 ako

$$CD_n^{(J)}(x) = \frac{1}{J-1}\sum_{j=d+1}^{J} CD_n^j(x).$$

V prípade konvexných obalov sa však aj členy CD^j a CD_n^j pre $j > d+1$ môžu javiť ako zbytočné, pretože nesú do istej miery redundantnú informáciu, ktorá bola obsiahnutá už v $(d+1)$-tom člene. Ak napríklad pre jednoduchosť v dvoch rozmeroch pre prípad náhodného výberu o rozsahu $n \in \mathbb{N}$ leží bod vo všetkých simplexoch, bude tiež ležať vo všetkých konvexných obaloch pre $j > (d+1)$. Ak naopak neleží v žiadnom simplexe, rovnako nebude ležať v žiadnom konvexnom obale pre viac bodov. Podobná úvaha vedie k domnienke, že ak bod leží v $m \in \mathbb{N}$ simplexoch, bude ležať v presne $m(n-3)$ konvexných obaloch tvorených štyrmi bodmi. Takto jednoducho však výpočet determinovaný nie je, pretože niektoré štvorbodové kombinácie sa môžu opakovať. Vidíme však, že všetky členy pre $j > (d+1)$ sú na rozdiel od pásovej hĺbky do istej miery určené členom $j = d+1$. Tejto myšlienke sa budeme v jednoduchšom prípade viac venovať v kapitole 6.

Rozšírenie zo simplexov na konvexné obaly získava na dôležitosti v prípade, ak nepočítame hĺbku iba jedného bodu, ale počítame súčasne hĺbku celej množiny bodov, ktorá môže byť až nekonečná tak, ako vo funkcionálnom prípade v časti 3.3. Takáto množina bodov nemusí byť celá obsiahnutá ani v jednom simplexe, ale môže byť obsiahnutá v nejakom väčšom konvexnom obale. Preto môžu v tomto prípade hrať rolu aj pásové (resp. konvexné) hĺbky pre $J > (d+1)$.

V jednoduchom prípade, ak počítame hĺbku jediného bodu, sme došli k záveru, že afinne invariantná pásová hĺbka nie je nič iné ako nejaké zovšeobecnenie simplexovej hĺbky dát. Ak budeme teda vyžadovať od pásovej hĺbky pre konečnorozmerné dáta vlastnosť afinnej invariancie, budeme pre jednobodové množiny používať simplexovú hĺbku a pre množiny s veľkou kardinalitou jej zovšeobecnenie, konvexnú hĺbku. K týmto úvahám sa neskôr vrátime v kapitole 4.

Jeden dôležitý rozdiel medzi konvexnou, prípadne simplexovou hĺbkou a pásovou hĺbkou je ten, že konvexná (resp. simplexová) hĺbka sa nedá zovšeobecňovať v zmysle

3.7 a 3.8. Nevieme totiž pre konvexné obaly nájsť analógiu toho, v koľkých súradniciach bod leží v páse. Iným spôsobom však toto zovšeobecnenie použijeme v prípade funkcionálnych dát, ako uvidíme neskôr v kapitole 4.

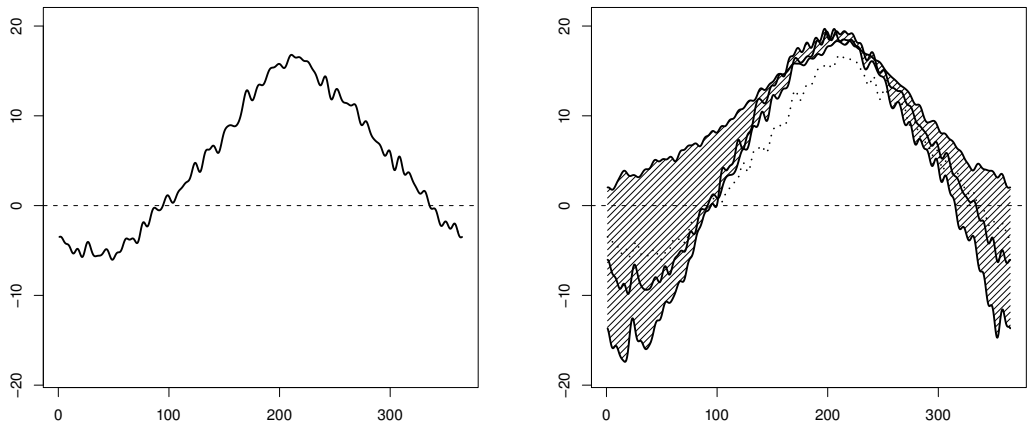
Pristúpme teraz k rozšíreniu konečnorozmerných pásových hĺbok na prípad funkcionálnych dát.

3.3 Pásové hĺbky pre funkcionálne dáta

Ako sme naznačili v diskusii o zobrazovaní bodov v paralelných súradniciach na strane 26, je možné zaviesť pásové hĺbky aj pre rozdelenia na priestoroch funkcií definovaných na uzavretom intervale. Opäť sa budeme obmedzovať na priestor $C([0,1])$ spojitých funkcií na uzavretom intervale $[0,1] \subset \mathbb{R}$. Zaveďme preto najprv značenie, ktoré budeme používať. Ako *graf funkcie* $x \in C([0,1])$ budeme označovať množinu

$$G(x) = \left\{ (t, x(t))^T : t \in [0,1] \right\} \subset \mathbb{R}^2,$$

ktorá je vlastne nespočetnorozmernou analógiou d-bodovej množiny vektoru v paralelných súradniciach vykreslených v $\mathbb{R}^2$. Graf funkcie môžeme vidieť v prvej časti obrázku 3.9. Definujme ďalej rovnako ako v konečnorozmernom prípade pás tvorený



Obr. 3.9: Funkcia a funkcionálny pás.

množinou funkcií.

Definícia:(Pás v $C([0,1])$)
Pre funkcie $x_1, \ldots, x_n \in C([0,1])$ definujeme ako

$$B(x_1, \ldots, x_n) = \left\{ (t,y)^T : t \in [0,1], \min_{i=1,\ldots,n} x_i(t) \leq y \leq \max_{i=1,\ldots,n} x_i(t) \right\}.$$

pás tvorený funkciami $x_1, \ldots, x_n$.

Pás je vlastne uzavretá množina v $\mathbb{R}^2$ ohraničená minimom a maximom funkcií. V druhej časti obrázku 3.9 vidíme pás tvorený tromi funkciami ako šrafovanú množinu, pričom funkcia z ľavej časti obrázka leží v tomto páse pre body $t \in [0,100] \cup [250,365]$.

Teraz môžeme zaviesť pásovú hĺbku pre funkcionálne dáta tak ako López-Pintado a Romo [15, 16] analogicky konečnorozmernému prípadu.

Definícia:(Pásová hĺbka v $C([0,1])$-populačná verzia)
Nech $x \in C([0,1])$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}(C([0,1]))$ a $X_1,\ldots,X_J$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$LP^j(x;P) = P\left(G(x) \subseteq B\left(X_1,\ldots,X_j\right)\right). \tag{3.21}$$

Potom *pásovú hĺbku J-teho rádu funkcie x vzhľadom k rozdeleniu pravdepodobnosti P* definujeme ako

$$LP^{(J)}(x;P) = \frac{1}{J-1}\sum_{j=2}^{J} LP^j(x;P).$$

Člen $LP^j(x;P)$ interpretujeme ako pravdepodobnosť, že funkcia x bude celým svojím grafom ležať v páse danom nejakým náhodným výberom o rozsahu j z rozdelenia P.

Výberovú verziu zavedieme pomocou U-štatistiky príslušnej 3.21.

Definícia:(Pásová hĺbka v $C([0,1])$-výberová verzia)
Nech $x \in C([0,1])$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}(C([0,1]))$, $n \geq J$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$LP_n^j(x;\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \mathbb{I}\left[G(x) \subseteq B\left(X_{i_1},\ldots,X_{i_j}\right)\right]. \tag{3.22}$$

Potom *pásovú hĺbku J-teho rádu funkcie x vzhľadom k náhodnému výberu* $\mathbb{X}$ definujeme ako

$$LP_n^{(J)}(x;\mathbb{X}) = \frac{1}{J-1}\sum_{j=2}^{J} LP_n^j(x;\mathbb{X}).$$

Parameter J v oboch definíciách určuje maximálny počet funkcií použitých na konštrukciu pásu. Tento koeficient sa zvyčajne pre zjednodušenie výpočtu rovnako ako pre konečnorozmerné pásové hĺbky volí $J = 2$ alebo 3. S rastúcim J sa ale $LP_n^{(J)}(x)$ už príliš nemení, ako poznamenali na základe simulácií López-Pintado a Romo [16].

Funkcia x bude mať nulovú výberovú pásovú hĺbku práve vtedy, keď ľubovoľne malá časť jej grafu bude ležať mimo pásu $B(X_1,\ldots,X_n)$. Takáto podmienka nenulovosti hĺbky funkcie sa však v prípade konečného náhodného výberu z nespočetnorozmerného rozdelenia pravdepodobnosti P ukazuje často ako príliš reštriktívna. K zovšeobecneniu pásovej hĺbky pre funkcie preto použijeme myšlienku zovšeobecnených konečnorozmerných pásových hĺbok. Pretože

$$LP^j(x;P) = P\left(G(x) \subseteq B\left(X_1,\ldots,X_j\right)\right) = \mathrm{E}\left[\mathbb{I}\left[G(x) \subseteq B\left(X_1,\ldots,X_j\right)\right]\right],$$

môžeme v definícii 3.21 namiesto indikátoru použiť inú, menej reštriktívnu mieru toho, že funkcia x neleží dostatočne v páse definovanom j funkciami náhodného výberu. Prirodzenou možnosťou je použitie Lebesgueovej miery λ na intervale $[0,1]$ tých bodov, ktoré ležia v páse. Takýmto spôsobom sa zavádza zovšeobecnená pásová hĺbka.

Definícia:(Zovšeobecnená pásová hĺbka v $C([0,1])$-populačná verzia)
Nech $x \in C([0,1])$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}(C([0,1]))$ a $X_1,\ldots,X_J$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$ ako

$$A_j\left(x;X_1,\ldots,X_j\right) = \left\{t \in [0,1] : \min_{r=1,\ldots,j} X_r(t) \leq x(t) \leq \max_{r=1,\ldots,j} X_r(t)\right\}$$

množinu takých bodov $t \in [0,1]$, v ktorých funkcia x leží v páse tvorenom funkciami $X_1,\ldots,X_j$. Označme ďalej ako

$$GLP^j(x;P) = \mathrm{E}\left[\lambda\left(A_j\left(x;X_1,\ldots,X_j\right)\right)\right].$$

Potom *zovšeobecnená pásová hĺbka J-teho rádu funkcie x vzhľadom k rozdeleniu pravdepodobnosti P* je

$$GLP^{(J)}(x;P) = \frac{1}{J-1}\sum_{j=2}^{J} GLP^j(x;P).$$

Na zavedenie výberovej verzie znovu použijeme U-štatistiku.

Definícia:(Zovšeobecnená pásová hĺbka v $C([0,1])$-výberová verzia)
Nech $x \in C([0,1])$, $J = 2,3,\ldots$, nech $P \in \mathcal{P}(C([0,1]))$, $n \geq J$ a $\mathbb{X} = (X_1,\ldots,X_n)^T$ je náhodný výber z rozdelenia P. Označme pre $j = 2,\ldots,J$

$$GLP_n^j(x;\mathbb{X}) = \binom{n}{j}^{-1} \sum_{1\leq i_1<\ldots<i_j\leq n} \lambda\left(A_j\left(x;X_{i_1},\ldots,X_{i_j}\right)\right).$$

Potom *zovšeobecnená pásová hĺbka J-teho rádu funkcie x vzhľadom k náhodnému výberu* $\mathbb{X}$ je

$$GLP_n^{(J)}(x;\mathbb{X}) = \frac{1}{J-1}\sum_{j=2}^{J} GLP_n^j(x;\mathbb{X}). \tag{3.23}$$

Vlastnosti zovšeobecnenej pásovej hĺbky budú do veľkej miery podobné vlastnostiam pásovej hĺbky. Samozrejme je možné pásové hĺbky analogicky zovšeobecniť ďalej pre akúkoľvek voľbu miery náležania grafu funkcie do pásu.

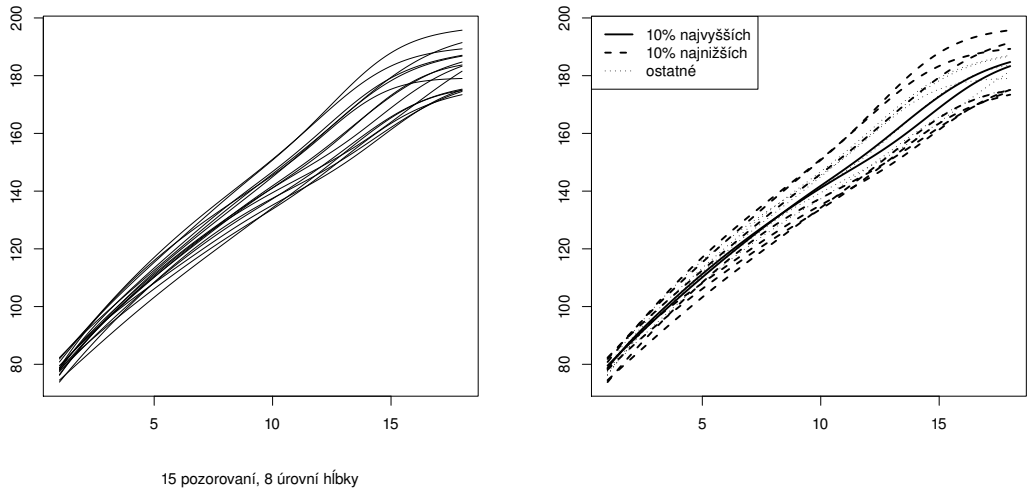
Fraimanova-Munizovej simplexová hĺbka tak ako sme ju definovali v 3.1 je špeciálnym prípadom zovšeobecnenej pásovej hĺbky rádu $J = 2$.

Ukážme teraz niekoľko príkladov použitia pásových hĺbok na funkcionálne dáta a porovnajme ich s indukovanými hĺbkami. Pretože zovšeobecnená pásová hĺbka rádu $J = 2$ aj 3 je ekvivalentná Fraimanovej-Munizovej simplexovej hĺbke, príklady jej použitia sme už videli v príkladoch 8, 9 a 10. Uvedieme teda iba niekoľko príkladov použitia pásovej hĺbky.

Príklad 12. Majme náhodný výber 15 detských rastových funkcií ako v príklade 9. Výsledky pri použití pásovej hĺbky pre odporúčanú voľbu $J = 3$ sú znázornené na obrázku 3.10. Z definície pásovej hĺbky vyplýva, že všetky funkcie tvoriace obal (t.j. minimum a maximum funkcií) náhodného výberu dostávajú triviálne nulovú hĺbku. Aj to je dôvodom, prečo boli funkcie pomerne slabo rozlíšené iba na 8 úrovní hĺbky. Ďalším dôvodom slabého rozlíšenia pozorovaní je, že pásová hĺbka môže nadobúdať iba konečne veľa hodnôt z množiny

$$\left\{\frac{1}{J-1}\sum_{j=2}^{J}\frac{k_j}{\binom{n}{j}} : k_j \in \left\{0,\ldots,\binom{n}{j}\right\}\right\}.$$

Okrem vysokého počtu funkcií s nulovou hĺbkou sa však výsledky zdajú byť pomerne uspokojivé, najhlbšie funkcie sú pozorovania ktoré môžeme považovať za typické voči náhodnému výberu, funkcie s malou hĺbkou sa môžu javiť aj ako odľahlé.

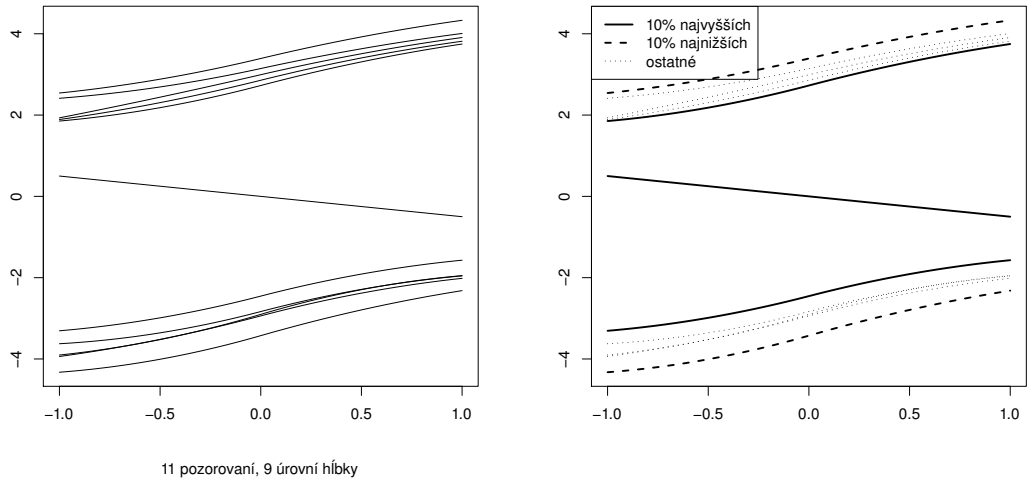


Obr. 3.10: Pásová hĺbka rádu 3 a rast chlapcov.

Príklad 13. Majme rovnaký náhodný výber simulovaných dát 10 rastúcich a jednej klesajúcej funkcie ako v príkladoch 6 a 10. Použime pásovú hĺbku rádu $J = 3$ (obrázok 3.11). Vidíme, že výsledky sú vizuálne totožné tým z príkladu 10. Vo výbere sú dve funkcie s nulovou hĺbkou, pretože obal pozorovaní je tvorený iba dvomi funkciami. Pretože funkcie náhodného výberu sa navzájom takmer nekrížia (majú málo bodov grafu spoločných pre viac funkcií), znamená to, že zovšeobecnená pásová hĺbka (Fraimanova-Munizovej hĺbka) a pásová hĺbka dávajú veľmi podobné výsledky. To vidíme aj v prípade funkcií s najvyššími hodnotami hĺbky.

Ako funkcie s 10 % najvyššími hodnotami hĺbky sú v oboch prípadoch rovnako ako v prípade Fraimanovych-Munizovej hĺbok označené tri funkcie ležiace „uprostred“ náhodného výberu. To je zrejmé, pretože pri výpočte hĺbok sa stále kladie dôraz na to, aby graf funkcie, ktorej hĺbku počítame, ležal čo najviac v páse tvorenom grafmi jednotlivých skupín z funkcií náhodného výberu. V oboch prípadoch prostredná, klesajúca funkcia, leží v každom páse obsahujúcom aspoň jednu funkciu zo zhluku vrchných funkcií a aspoň jednu funkciu zo zhluku spodných funkcií. Preto práve klesajúca funkcia opäť dostáva chybne najväčšiu hĺbku a stáva sa výberovou funkcionálnou analógiou mediánu.

Narazili sme na najväčšiu nevýhodu pásových hĺbok. Rovnako ako jednoduchá Fraimanova-Munizovej hĺbka totiž odlišujú dobre iba pozorovania odľahlé v polohe (pozorovanie odľahlé v polohe bude ležať v malom počte pásov), nie sú však schopné odlíšiť pozorovania odľahlé v tvare. Preto ak by sme aj v príklade 12 určovali pásovú hĺbku funkcie, ktorá sa drží blízko strednej hodnoty po celú dobu priebehu, ale fluktuuje (v krátkych intervaloch sa strieda v priebehu z pomaly klesajúcej na rýchlo rastúcu), dostala by tiež jednu z najvyšších hodnôt hĺbky. V prípade náhodného výberu veľmi hladkých striktne rastúcich funkcií však takáto divná funkcia nemôže byť označená ako typická iba preto že sa počas celého priebehu drží z pohľadu suprémovej metriky blízko strednej hodnoty.



Obr. 3.11: Pásová hĺbka rádu 3 a simulované dáta.

Ako sme práve ukázali, pásové hĺbky nie sú dobrým riešením, ak máme v náhodnom výbere aj funkcie odľahlé v tvare. Tento problém by sa však dal vyriešiť tým, že by sme do výpočtu hĺbky pridali informáciu o tvare funkcií. Týmto sa budeme zaoberať neskôr v kapitole 4.

Pozrime sa teraz na niekoľko základných vlastností pásových hĺbok.

Rovnako ako konečnorozmerné pásové hĺbky, aj funkcionálne pásové hĺbky sú vo výberovom prípade založené na koncepte U-štatistík. Pre konečnorozmerný prípad sme toto pozorovanie sformulovali ako tvrdenie 3.2, sformulujme ho teda pre úplnosť aj vo funkcionálnom prípade.

Tvrdenie 3.5. *Nech* $P \in \mathcal{P}(C([0,1]))$, $J = 2,3,\ldots$. *Potom* $LP_n^{(J)}$ *aj* $GLP_n^{(J)}$ *sú U-štatistiky rádu* J.

Dôkaz. Analogicky dôkazu tvrdenia 3.2. □

Nasleduje tvrdenie o tom, nakoľko funkcionálne pásové hĺbky splňujú podmienky kladené na štatistickú hĺbkovú funkciu. Tieto podmienky by mali byť analógiou podmienok kladených na štatistickú hĺbkovú funkciu v konečnorozmernom prípade tak, ako sme ich vyšetrovali pre indukované hĺbky v tvrdení 2.1.

Tvrdenie 3.6. *Nech* $P \in \mathcal{P}(C([0,1]))$ *a* $J = 2,3,\ldots$. *Potom platí:*

- *Nech* $x,a,b \in C([0,1])$, $a(t) \neq 0$ *pre každé* $t \in [0,1]$. *Potom*

$$LP^{(J)}(x;P_X) = LP^{(J)}(ax+b;P_{aX+b}).$$

- *Ak* $\mathbb{X}_n = (X_1,\ldots,X_n)^T$ *je náhodný výber z rozdelenia* P, *potom platí*

$$\sup_{\|x\|_\infty \geq M} LP^{(J)}(x;P) \xrightarrow[M\to\infty]{} 0,$$

$$\sup_{\|x\|_\infty \geq M} LP_n^{(J)}(x;\mathbb{X}_n) \xrightarrow[M\to\infty]{s.i.} 0.$$

- *$LP^{(J)}$ je zhora polospojité zobrazenie na $C([0,1])$. Ak naviac rozdelenie pravdepodobnosti P má absolútne spojité marginálne rozdelenia, potom $LP^{(J)}$ je spojité zobrazenie na $C([0,1])$.*

Dôkaz. Pozri López-Pintado a Romo [16]. □

Pre funkcionálnu pásovú hĺbku teda platí nekonečnorozmerná analógia afinnej invariancie (podmienka P1) aj podmienka nulovosti limity pre funkciu rastúcu v suprémovej norme nad všetky medze (podmienka P4). Platnosť podmienok P2 a P3 vyšetríme vo všeobecnejšom prípade v kapitole 4.

Pre rovnako spojitú množinu funkcií je možné odvodiť konzistenciu výberových pásových hĺbok voči populačnej verzii.

Tvrdenie 3.7. *Nech $P \in \mathcal{P}(C([0,1]))$ s absolútne spojitými marginálnymi rozdeleniami a $J = 2,3,\ldots$. Potom*

- *$LP_n^{(J)}$ rovnomerne konverguje k $LP^{(J)}$ na každej množine rovnako spojitých funkcií E, t.j.*

$$\sup_{x\in E}\left|LP_n^{(J)}(x) - LP^{(J)}(x)\right| \xrightarrow[n\to\infty]{s.i.} 0.$$

- *Nech $LP^{(J)}$ má jediné maximum v bode $m \in E$ a nech m_{nJ} je postupnosť funkcií v E taká, že $LP_n^{(J)}(m_{nJ}) = \sup_{x\in E} LP_n^{(J)}(x)$. Potom*

$$m_{nJ} \xrightarrow[n\to\infty]{s.i.} m$$

pre $n \to \infty$.

Dôkaz. Pozri López-Pintado a Romo [16]. □

Napriek tomu, že pásové hĺbky majú množstvo vlastností aké od dobrej hĺbkovej funkcie očakávame, v príklade 13 sme ukázali, že nie sú schopné odlišovať pozorovania odľahlé v tvare. Tento problém je možné odstrániť tým, že vo výpočte pásovej hĺbky použijeme diferencovateľnosť funkcií, teda čiste funkcionálnu vlastnosť, ktorá rozšíri možnosti inferencie aj na rozhodovanie o tom, či funkcia, ktorej hĺbku počítame, je tvarovo podobná funkciám náhodného výberu.

Kapitola 4

Geometricko-funkcionálna hĺbka

Problém pásovej hĺbky s rozoznávaním tvaru funkcií môžeme odstrániť, ak pripustíme diferencovateľnosť všetkých funkcií tvoriacich nosič rozdelenia P. Tým sa obmedzíme z priestoru spojitých funkcií na $[0,1]$ na priestor dostatočne hladkých funkcií na $[0,1]$. V takom prípade už budeme môcť pracovať nielen so samotnou funkčnou hodnotou pozorovania, ale aj s deriváciami niekoľkých rádov, čo podstatne rozšíri možnosti detekcie pozorovaní odľahlých v tvare.

Pri zavedení novej hĺbky funkcionálnych dát budeme postupovať v smere zovšeobecnenia ako zovšeobecnených pásových hĺbok, tak aj Fraimanovej-Munizovej hĺbky na funkcie spojite diferencovateľné do rádu $K \in \mathbb{N}_0$. Dokážeme niekoľko dôležitých vlastností novej hĺbky a na záver ilustrujeme jej použitie na predchádzajúce príklady.

Zaveďme však najprv značenie.

V celom ďalšom texte bude vždy $K \in \mathbb{N}_0$ označovať rád diferencovateľnosti všetkých uvažovaných funkcií a

$$\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_K)^T \in [0,\infty)^{K+1}$$

bude označovať vektor váh, teda napospol nenulových nezáporných konštánt ($\alpha \neq 0$, kde znakom 0 budeme označovať ako číslo $0 \in \mathbb{R}$, tak aj vektor $0_d \in \mathbb{R}^d$ pre každé $d \in \mathbb{N}$).

Urobme najprv dohovor o tom, čo budeme považovať za priestor $C^{(K)}([0,1])$. Pretože v ďalšom budeme potrebovať, aby k-ta derivácia funkcie $x \in C^{(K)}([0,1])$ bola spojitá na celom intervale $[0,1]$, budeme za k-tu deriváciu funkcie x v krajnom bode intervalu $[0,1]$ považovať jednostrannú deriváciu v tomto bode pre každé $k = 1, \ldots, K$. Preto môžeme k-tu deriváciu funkcie x považovať za funkciu spojitú na celom uzavretom intervale $[0,1]$.

Ďalej nech $x \in C^{(K)}([0,1])$ a $t \in [0,1]$. Zaveďme značenie pre vektor prvých $k = 0, \ldots, K$ derivácií funkcie x v bode t

$$x^{(0,\ldots,k)}(t) = \left(x^{(0)}(t), x^{(1)}(t), \ldots, x^{(k)}(t)\right)^T \in \mathbb{R}^{k+1}.$$

Ďalej definujme *K-rozmerný simplex tvorený funkciami* $\{g_i\}_{i=1}^{K+2} \subset C^{(K)}([0,1])$ *v bode* $t \in [0,1]$ ako simplex v $K+1$-rozmernom euklidovskom priestore tvorený bodmi

$$\left\{g_i^{(0,\ldots,K)}(t)\right\}_{i=1}^{K+2} \subset \mathbb{R}^{K+1}$$

a označme ho $\mathbb{S}_{g_1,\ldots,g_{K+2}}(t)$. Poznamenajme na tomto mieste ešte raz, že za simplex považujeme aj degenerovaný prípad, ak body $\left\{g_i^{(0,\ldots,K)}(t)\right\}_{i=1}^{K+2}$ sú lineárne závislé.

Za normu v priestore $C^{(K)}([0,1])$ berieme prirodzenú suprémovú normu pre funkcie spojite diferencovateľné do rádu K

$$\|x\|^{(K)} = \max_{k=0,\ldots,K} \sup_{t\in[0,1]} \left|x^{(k)}(t)\right|. \tag{4.1}$$

4.1 K-pásová hĺbka

Priamočiarym zovšeobecnením pásových hĺbok tak, aby odlišovali pozorovania odľahlé v tvare, by bola jednoduchá hĺbka súčtového typu

$$D(x;P) = \frac{1}{K}\sum_{k=0}^{K} LP^{(J)}\left(x^{(k)};P^{(k)}\right), \tag{4.2}$$

prípadne súčinového typu

$$D(x;P) = \left(\prod_{k=0}^{K} LP^{(J)}\left(x^{(k)};P^{(k)}\right)\right)^{\frac{1}{K}}, \tag{4.3}$$

kde $x^{(k)}$ je k-ta derivácia funkcie $x \in C^{(K)}([0,1])$ a $P^{(k)}$ označuje rozdelenie pravdepodobnosti k-tej derivácie funkcií z rozdelenia pravdepodobnosti P. Výberová verzia by sa dodefinovala analogicky pásovým hĺbkam. Takáto hĺbka by zrejme bola schopná odlišovať pozorovania odľahlé v tvare, pretože ak by sa funkcia x líšila tvarom od funkcií náhodného výberu (ako v prípade simulovaných funkcionálnych dát v príkladoch 6, 10 a 13), líšila by sa výrazne práve v niektorej derivácii a tak by mal príslušný sčítanec v prípade 4.2 alebo činiteľ v 4.3 nízku hodnotu, čo by znižovalo celkovú hĺbku funkcie. V prípade 4.3 by dokonca jeden nulový činiteľ mohol znížiť celkovú hĺbku funkcie x až na nulu, čím by sa pozorovanie stalo odľahlým iba preto, že je odľahlým v niektorej derivácii.

Na takéto jednoduché hĺbky by sa však neprenášali dobré vlastnosti pásových hĺbok ako (slabá) afinná invariancia, maximalita v bode stredu symetrie, prípadne konzistencia výberovej verzie. Ak však 4.2 trochu upravíme, ľahko sa presvedčíme, že všetky dobré vlastnosti zovšeobecnenej pásovej hĺbky zostanú zachované a navyše hĺbka bude schopná odlišovať pozorovania odľahlé v tvare.

Definujme teraz takéto rozšírenie zovšeobecnenej pásovej hĺbky, a ako uvidíme aj Fraimanovej-Munizovej hĺbky, na funkcie z $C^{(K)}([0,1])$ s pomocou konečnorozmernej simplexovej hĺbky.

Definícia:(K-pásová hĺbka-populačná verzia)
Nech $K \in \mathbb{N}_0$, $x \in C^{(K)}([0,1])$ a $X_1,\ldots,X_{K+2}$ je náhodný výber z $P \in \mathcal{P}\left(C^{(K)}([0,1])\right)$, $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_K) \in [0,\infty)^{K+1}$, $\alpha \neq 0$. Potom *K-pásovú hĺbku s váhami* α *funkcie x vzhľadom k P* definujeme ako

$$KLP^{(K)}(x;P) = \frac{\sum_{k=0}^{K} KLP^{k}(x;P)\,\alpha_k}{\sum_{i=0}^{K}\alpha_i},$$

kde

$$KLP^{k}(x;P) = \int_0^1 P\left(x^{(0,\ldots,k)}(t) \in \mathbb{S}_{X_1,\ldots,X_{k+2}}(t)\right)\mathrm{d}t.$$

Na výberovú verziu K-pásovej hĺbky sa môžeme pozerať aj ako na populačnú verziu K-pásovej hĺbky voči empirickému rozdeleniu náhodného výberu.

Definícia:(K-pásová hĺbka-výberová verzia)
Nech $K \in \mathbb{N}_0$, $n \in \mathbb{N}$, $n \geq (K+2)$, $x \in C^{(K)}([0,1])$, $\mathbb{X} = (X_1, \ldots, X_n)^T$ je náhodný výber z $P \in \mathcal{P}\left(C^{(K)}([0,1])\right)$, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_K) \in [0,\infty)^{K+1}$, $\alpha \neq 0$. Potom *K-pásovú hĺbku s váhami α funkcie x vzhľadom k náhodnému výberu $\mathbb{X}$* definujeme ako

$$KLP_n^{(K)}(x;\mathbb{X}) = \frac{\sum_{k=0}^{K} KLP_n^k(x;\mathbb{X})\,\alpha_k}{\sum_{i=0}^{K}\alpha_i},$$

kde

$$KLP_n^k(x;\mathbb{X}) = \binom{n}{k+2}^{-1} \sum_{1\leq i_1<\ldots<i_{k+2}\leq n} \lambda\left(t \in [0,1] : x^{(0,\ldots,k)}(t) \in \mathbb{S}_{X_{i_1},\ldots,X_{i_{k+2}}}(t)\right). \quad (4.4)$$

Sčítanec KLP_n^k nazvime *k-ty sčítanec K-pásovej hĺbky* pre $k = 0, \ldots, K$.

Takto definovaný funkcionál môžeme chápať ako vážený aritmetický priemer funkcionálov KLP_n^k s váhami α. Zovšeobecnená pásová hĺbka rádu $J = 2$, rovnako ako aj Fraimanova-Munizovej hĺbka sú zrejme špeciálnym prípadom K-pásovej hĺbky pre $K = 0$ a $\alpha_0 = 1$.

Ako v celom texte, v prípade, že bude zrejmé, voči akému rozdeleniu sa K-pásová hĺbka počíta, budeme v značení vynechávať argument pravdepodobnostného rozdelenia. Dokážme teraz niektoré základné vlastnosti novej hĺbky.

Tvrdenie 4.1. *Nech $K \in \mathbb{N}_0$ a $P \in \mathcal{P}\left(C^{(K)}([0,1])\right)$. Potom $KLP_n^{(K)}$ je U-štatistika rádu $(K+2)$.*

Dôkaz. Analogicky dôkazu tvrdenia 3.2. □

Na k-ty sčítanec K-pásovej hĺbky sa môžeme pozerať až na multiplikatívnu konštantu ako na integrál z $(k+1)$-rozmernej simplexovej hĺbky voči združenému rozdeleniu funkčnej hodnoty a prvých k derivácií funkcií náhodného výberu v bode $t \in [0,1]$. Takáto reprezentácia sa bude hodiť pri ďalších úvahách. Označme preto pre $k = 0, \ldots, K$ a pre $t \in [0,1]$ marginálne rozdelenie derivácií rádov $\{0, \ldots, k\}$ funkcie z rozdelenia pravdepodobnosti P na $C^{(K)}([0,1])$ v bode $t \in [0,1]$ ako $P_t^{(0,\ldots,k)}$, resp. pre zdôraznenie k akej náhodnej veličine X sa marginálne rozdelenie vzťahuje ako $P_{X(t)}^{(0,\ldots,k)}$.

Tvrdenie 4.2. *Nech $K \in \mathbb{N}_0$ a P_X je rozdelenie pravdepodobnosti náhodnej veličiny X na $C^{(K)}([0,1])$, nech $a, b \in C^{(K)}([0,1])$, $a(t) \neq 0$ $[\lambda]$-skoro všade. Potom pre každú funkciu $x \in C^{(K)}([0,1])$ platí rovnosť*

$$KLP^{(K)}(x;P_X) = KLP^{(K)}(ax+b;P_{aX+b}).$$

Dôkaz. Využijeme reprezentáciu K-pásovej hĺbky ako integrálu z k-rozmernej simplexovej hĺbky pre $k = 1,\ldots,K+1$. Vezmime l'ubovol'né pevné $t \in [0,1]$ a pevné $k = 0,\ldots,K$. Podl'a Leibnizovej formuly pre výpočet derivácie súčinu funkcií platí

$$(ax+b)^{(k)}(t) = b^{(k)}(t) + \sum_{i=0}^{k} \binom{k}{i} x^{(i)}(t)\, a^{(k-i)}(t).$$

Pozrime sa teraz na to, ako bude pre $x \in C^{(K)}([0,1])$ po afinnej transformácii vyzerať obraz bodu $x^{(0,\ldots,k)}(t) \in \mathbb{R}^{k+1}$. Podl'a predchádzajúceho platí

$$\begin{aligned}(ax+b)^{(0,\ldots,k)}(t) &= \begin{pmatrix} a^{(0)}(t)x^{(0)}(t) + b^{(0)}(t) \\ a^{(1)}(t)x^{(0)}(t) + a^{(0)}(t)x^{(1)}(t) + b^{(1)}(t) \\ \ldots \\ \sum_{i=0}^{k}\binom{k}{i}x^{(i)}(t)\,a^{(k-i)}(t) + b^{(k)}(t) \end{pmatrix} \\ &= Ax^{(0,\ldots,k)}(t) + b^{(0,\ldots,k)}(t),\end{aligned}$$

kde sme označili ako

$$A = \begin{pmatrix} a^{(0)}(t) & 0 & 0 & \ldots & 0 \\ a^{(1)}(t) & a^{(0)}(t) & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \binom{k}{0}a^{(k)}(t) & \binom{k}{1}a^{(k-1)}(t) & \binom{k}{2}a^{(k-2)}(t) & \ldots & \binom{k}{k}a^{(0)}(t) \end{pmatrix}$$

maticu konštánt závislú na hodnotách vektoru $a^{(0,\ldots,k)}(t)$.

Vidíme, že sa jedná o afinnú transformáciu vektoru $x^{(0,\ldots,k)}(t)$ v priestore $\mathbb{R}^{k+1}$. Podl'a vety o afinnej invariancii k-rozmernej simplexovej hĺbky $SD(.,.)$ (tvrdenie 1.1) pre regulárnu maticu A platí

$$\begin{aligned}KLP^k(ax+b; P_{aX+b}) &= \int_0^1 SD\left(Ax^{(0,\ldots,k)}(t) + b^{(0,\ldots,k)}(t); P^{(0,\ldots,k)}_{(aX+b)(t)}\right) \mathrm{d}t \\ &= \int_0^1 SD\left(x^{(0,\ldots,k)}(t); P^{(0,\ldots,k)}_{X(t)}\right) \mathrm{d}t \\ &= KLP^k(x; P_X).\end{aligned} \tag{4.5}$$

Matica A je zrejme regulárna práve vtedy, ak $a^{(0)}(t) \neq 0$, čo je práve podmienka $a(t) \neq 0$. Pre platnosť 4.5 však postačuje afinná invariancia pre simplexovú hĺbku $[\lambda]$-s.v., preto postačuje $a \neq 0$ pre $[\lambda]$-s.v. $t \in [0,1]$. Sčítance K-pásovej hĺbky sú teda afinne invariantné, preto aj samotná K-pásová hĺbka je afinne invariantná. □

Poznamenajme, že celý dôkaz tvrdenia 4.2 zostáva v platnosti nielen pre hĺbku počítanú pomocou simplexov, teda konvexných obalov množín bodov v $\mathbb{R}^k$, ale bez zmeny by platilo aj silnejšie tvrdenie, v ktorom by sme počítali hĺbku funkcie pomocou konvexných obalov bodov nie nutne tvoriacich simplex. Preto afinná invariancia zostáva v platnosti aj pre K-pásovú hĺbku upravenú tak, aby v k-tom sčítanci zohľadňovala nielen to, či príslušný bod $x^{(0,\ldots,k)}(t)$ leží v $(k+1)$-rozmernom simplexe, ale aj to, či leží pre $j \in \mathbb{N}$ v konvexnom obale tvorenom j bodmi funkcií z náhodného výberu. V takomto prípade je zrejmé v akom zmysle je K-pásová hĺbka rozšírením zovšeobecnenej pásovej hĺbky J-teho rádu pre každé J.

Tvrdenie 4.3. *Nech $K \in \mathbb{N}_0$ a $P \in \mathcal{P}\left(C^{(K)}([0,1])\right)$. Potom $KLP^{(K)}$ je zhora polospojitý funkcionál na priestore $C^{(K)}([0,1])$.*

Dôkaz. Najprv ukážeme, že pre každé $k = 0,\ldots,K$,pre každé $t \in [0,1]$ a pre každú postupnosť funkcií $\{x_n\}_{n=1}^{\infty} \subset C^{(K)}$ platí

$$x_n \xrightarrow[n\to\infty]{} x \Rightarrow x_n^{(0,\ldots,k)}(t) \xrightarrow[n\to\infty]{} x^{(0,\ldots,k)}(t).$$

To platí, pretože

$$\begin{aligned}
x_n \xrightarrow[n\to\infty]{} x &\Leftrightarrow \|x_n - x\|^{(K)} \xrightarrow[n\to\infty]{} 0 \\
&\Rightarrow \max_{i=0,\ldots,k} \sup_{t\in[0,1]} \left|x_n^{(i)}(t) - x^{(i)}(t)\right| \xrightarrow[n\to\infty]{} 0 \\
&\Leftrightarrow \forall i = 0,\ldots,k : \sup_{t\in[0,1]} \left|x_n^{(i)}(t) - x^{(i)}(t)\right| \xrightarrow[n\to\infty]{} 0 \\
&\Leftrightarrow \forall i = 0,\ldots,k, \forall t \in [0,1] : \left|x_n^{(i)}(t) - x^{(i)}(t)\right| \xrightarrow[n\to\infty]{} 0 \qquad (4.6) \\
&\Leftrightarrow \forall t \in [0,1] : \sum_{i=0}^{k} \left(x_n^{(i)}(t) - x^{(i)}(t)\right)^2 \xrightarrow[n\to\infty]{} 0 \\
&\Leftrightarrow \forall t \in [0,1] : \left\|x_n^{(0,\ldots,k)}(t) - x_n^{(0,\ldots,k)}(t)\right\| \xrightarrow[n\to\infty]{} 0 \\
&\Leftrightarrow \forall t \in [0,1] : x_n^{(0,\ldots,k)}(t) \xrightarrow[n\to\infty]{} x^{(0,\ldots,k)}(t).
\end{aligned}$$

Označme ako

$$\mathbb{X} = (X_1,\ldots,X_{k+2})^T : \Omega \to \left(C^{(K)}([0,1])\right)^{k+2}$$

náhodný výber $(k+2)$ funkcií z rozdelenia P. Potom pre každý sčítanec K-pásovej hĺbky môžeme písať pre postupnosť funkcií $x_n \xrightarrow[n\to\infty]{} x$

$$\begin{aligned}
\limsup_{n\to\infty} KLP^k(x_n) &= \limsup_{n\to\infty} \int_0^1 P\left(x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}(t)\right) \mathrm{d}t \\
&= \limsup_{n\to\infty} \int_0^1 \int_\Omega \mathbb{I}\left[x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}(t)\right] \mathrm{d}P(\mathbb{X})\,\mathrm{d}t \\
&\le \int_0^1 \limsup_{n\to\infty} \int_\Omega \mathbb{I}\left[x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}(t)\right] \mathrm{d}P(\mathbb{X})\,\mathrm{d}t \\
&\le \int_0^1 \int_\Omega \limsup_{n\to\infty} \mathbb{I}\left[x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}(t)\right] \mathrm{d}P(\mathbb{X})\,\mathrm{d}t \\
&\le \int_0^1 \int_\Omega \mathbb{I}\left[x^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}(t)\right] \mathrm{d}P(\mathbb{X})\,\mathrm{d}t \\
&= KLP^k(x_n).
\end{aligned}$$

Pretože všetky funkcie vo výpočte sú nezáporné a rovnako obmedzené konštantnou funkciou 1, ktorá je integrovateľná na intervale $[0,1]$ podľa Lebesgueovej miery λ, rovnako ako aj na priestore Ω podľa pravdepodobnostnej miery P, mohli sme vo výpočte niekoľkokrát použiť Fatouovo lemma pre limes superior. Prvé dve nerovnosti vyplývajú práve z Fatouovho lemma a tretia nerovnosť vyplýva z toho, že indikátorová funkcia uzavretej množiny, akou zrejme simplex je, je zhora polospojitá. Týmto

sme teda dokázali, že sčítance K-pásovej hĺbky sú zhora polospojité funkcie. Nakoniec, rovnako s použitím Fatouovho lemma pre limes superior tentokrát pre konečný súčet funkcií $KLP^k(x_n)$ zhora obmedzených opäť konštantou 1 môžeme ukázať

$$\begin{aligned}\limsup_{n\to\infty} KLP^{(K)}(x_n) &= \limsup_{n\to\infty} \frac{\sum_{k=0}^{K} KLP^k(x_n)\,\alpha_k}{\sum_{i=0}^{K}\alpha_i} \\ &\leq \frac{1}{\sum_{i=0}^{K}\alpha_i}\sum_{k=0}^{K}\alpha_k \limsup_{n\to\infty} KLP^k(x_n) \\ &\leq \frac{1}{\sum_{i=0}^{K}\alpha_i}\sum_{k=0}^{K}\alpha_k KLP^k(x) \\ &= KLP^{(K)}(x).\end{aligned}$$

Týmto sme dokázali, že K-pásová hĺbka je zhora polospojitý funkcionál pre každé rozdelenie pravdepodobnosti P. □

Poznamenajme ešte, že v prípade, že by sme nedefinovali K-pásovú hĺbku pomocou miery náležania bodov do uzavretých simplexov, ale namiesto toho by sme za simplexy považovali otvorené množiny, platilo by, že K-pásová hĺbka je zdola polospojitý funkcionál a dôkaz by zostal rovnaký s jedinou modifikáciou, že namiesto Fatouovho lemma pre limes superior by sme použili Fatouovo lemma pre limes inferior.

V ďalšom budeme potrebovať analógiu absolútnej spojitosti rozdelenia pravdepodobnosti P na nejakej triede hladkých funkcií $E \subset C^{(K)}([0,1])$, kde pripúšťame aj možnosť $E \equiv C^{(K)}([0,1])$. Tú zavedieme pomocou absolútnej spojitosti marginálnych rozdelení pre $[\lambda]$-s.v. $t \in [0,1]$.

Definícia:(absolútne spojité rozdelenie na $C^{(K)}([0,1])$)
Rozdelenie pravdepodobnosti P na $E \subset C^{(K)}([0,1])$ je *absolútne spojité*, ak platí pre $[\lambda]$-s.v. $t \in [0,1]$, že marginálne rozdelenie $P_t^{(0,\dots,K)}$ je absolútne spojité rozdelenie na priestore $\mathbb{R}^{K+1}$.

Za predpokladu absolútnej spojitosti rozdelenia platí silnejšie tvrdenie ako tvrdenie 4.3 o spojitosti K-pásovej hĺbky.

Tvrdenie 4.4. *Nech $K \in \mathbb{N}_0$ a P je absolútne spojité rozdelenie pravdepodobnosti na $C^{(K)}([0,1])$. Potom $KLP^{(K)}$ je spojitý funkcionál na priestore $C^{(K)}([0,1])$.*

Dôkaz. Dôkaz tvrdenia založíme na vete o spojitosti integrálu závislého na parametre, ktorú je možné nájsť napríklad v knihe Lukeša a Malého [17]. Zrejme stačí dokázať spojitosť sčítancov K-pásovej hĺbky KLP^k pre každé $k = 0,\dots,K$. Majme teda pevné, ale ľubovoľné $k \in \{0,\dots,K\}$. Máme

$$\begin{aligned}&KLP^k : C^{(K)} \to [0,1], \\ &KLP^k : x \mapsto \int_0^1 SD\left(x^{(0,\dots,k)}(t), P_t^{(0,\dots,k)}\right) \mathrm{d}t.\end{aligned}$$

Definujme funkciu Ψ dvoch premenných, a to jednej funkcionálnej a druhej skalárnej ako

$$\begin{aligned}\Psi : C^{(K)}([0,1]) \times [0,1] &\to [0,1], \\ \Psi(x,t) &= SD\left(x^{(0,\dots,k)}(t); P_t^{(0,\dots,k)}\right).\end{aligned}$$

Potom je

$$KLP^k : x \mapsto \int_0^1 \Psi(x,t)\,\mathrm{d}t,$$

a na takéto zobrazenie aplikujeme vetu o spojitosti integrálu závislého na parametre. K tomu potrebujeme overiť tri predpoklady vety, a to:

Spo1 Pre všetky $x \in C^{(K)}([0,1])$ a pre $[\lambda]$-skoro všetky $t \in [0,1]$ je funkcia $\Psi(.,t)$ spojitá v x.

Spo2 Pre všetky $x \in C^{(K)}([0,1])$ je funkcia $\Psi(x,.)$ $[\lambda]$-merateľná.

Spo3 Existuje $[\lambda]$-integrovateľná funkcia g na $[0,1]$ taká, že pre všetky $x \in C^{(K)}([0,1])$ a pre všetky $t \in [0,1]$ je $|\Psi(x,t)| \leq g(t)$.

Zrejme platí $|\Psi(x,t)| \leq 1$ a konštantná funkcia je na konečnom intervale integrovateľná, z čoho plynie platnosť podmienky Spo3. Ďalej, podmienka merateľnosti Spo2 je zrejme tiež triviálne splnená. Dokážme teraz podmienku spojitosti Spo1. Nech $t \in [0,1]$ je pevné. Ukážeme, že platí

$$x_n \xrightarrow[n\to\infty]{} x \Rightarrow \Psi(x_n,t) \xrightarrow[n\to\infty]{} \Psi(x,t).$$

Nech $x_n \xrightarrow[n\to\infty]{} x$. Potom platí $x_n^{(0,\ldots,k)}(t) \xrightarrow[n\to\infty]{} x^{(0,\ldots,k)}(t)$ pre každé $t \in [0,1]$ a pre každé $k = 0,\ldots,K$, ako sme už ukázali v 4.6. Nech $k = 1$. Ako uvidíme, dôkaz pre iné hodnoty k by sa robil úplne analogicky. Označme ako

$$\mathbb{X} = (X_1, X_2, X_3)^T : \Omega \to \mathbb{R}^2$$

náhodný výber troch dvojrozmerných náhodných vektorov z absolútne spojitého marginálneho rozdelenia $P_t^{(0,1)}$. Platí, že

$$\begin{aligned}
|\Psi(x_n,t) - \Psi(x,t)| &= \left| SD\left(x_n^{(0,\ldots,k)}(t); P_t^{(0,1)}\right) - SD\left(x^{(0,\ldots,k)}(t); P_t^{(0,1)}\right)\right| \\
&= \left| P_t^{(0,1)}\left(x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right) - P_t^{(0,1)}\left(x^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right)\right| \\
&= \left| \int_\Omega \mathbb{I}\left[x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right] - \mathbb{I}\left[x^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right] \mathrm{d}P_t^{(0,1)}(\mathbb{X})\right| \\
&\leq \int_\Omega \left|\mathbb{I}\left[x_n^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right] - \mathbb{I}\left[x^{(0,\ldots,k)}(t) \in \mathbb{S}_{\mathbb{X}}\right]\right| \mathrm{d}P_t^{(0,1)}(\mathbb{X}) \\
&\leq 3P_t^{(0,1)}(A_n),
\end{aligned} \tag{4.7}$$

kde

$$A_n = \left\{ (X_1,X_2)^T : \overline{X_1X_2} \cap \overline{x_n^{(0,\ldots,k)}(t)\, x^{(0,\ldots,k)}(t)} \neq \emptyset \right\},$$

takže množina takých dvojíc bodov tvoriacich vrcholy náhodného simplexu, že úsečka nimi tvorená má spoločný bod s úsečkou tvorenou bodmi $x_n^{(0,\ldots,k)}(t)$ a $x^{(0,\ldots,k)}(t)$. Podľa Fatouovho lemma pre limes superior platí

$$\limsup_{n\to\infty} P_t^{(0,1)}(A_n) \leq P_t^{(0,1)}\left(\limsup_{n\to\infty} A_n\right). \tag{4.8}$$

Limes superior v nerovnosti 4.8 môžeme rozumieť ako vo funkcionálnom zmysle (ako limes superior indikátorových funkcií $\mathbb{I}\left[(X_1,X_2)^T \in A_n\right]$), tak aj v množinovom zmysle (ako množinové limes superior postupnosti množín A_n). To plynie z ekvivalencie

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{I}\left[(X_1,X_2)^T \in A_n\right] = 1 &\Leftrightarrow \forall l \in \mathbb{N}\ \exists n \geq l:\ (X_1,X_2)^T \in A_n \\ &\Leftrightarrow (X_1,X_2)^T \in \bigcap_{l=1}^{\infty}\bigcup_{n=l}^{\infty} A_n \\ &\Leftrightarrow (X_1,X_2)^T \in \limsup_{n\to\infty} A_n.\end{aligned}$$

Pretože $x_n^{(0,\ldots,k)}(t) \xrightarrow[n\to\infty]{} x^{(0,\ldots,k)}(t)$ je podľa 4.6 pre množinové limes superior

$$\limsup_{n\to\infty} A_n = \left\{(X_1,X_2)^T : x^{(0,\ldots,k)}(t) \in \overline{X_1X_2}\right\}$$

a tak z absolútnej spojitosti marginálneho rozdelenia $P_t^{(0,1)}$ s použitím 4.8 dostávame

$$0 \leq \lim_{n\to\infty} P_t^{(0,1)}(A_n) \leq P_t^{(0,1)}\left(\limsup_{n\to\infty} A_n\right) = 0.$$

Podľa 4.7 je konečne $\Psi(.,t)$ spojitá v x a tým dostávame podľa vety o spojitosti integrálu závislého na parametre spojitosť KLP^k na $C^{(K)}([0,1])$. □

Pre absolútne spojité angulárne symetrické rozdelenia na $C^{(K)}([0,1])$ je K-pásová hĺbka je maximalizovaná vo funkcii tvoriacej stred symetrie rozdelenia.

Tvrdenie 4.5. *Nech $K \in \mathbb{N}_0$. Nech P je absolútne spojité angulárne symetrické rozdelenie na $C^{(K)}([0,1])$ so stredom symetrie vo funkcii $\theta \in C^{(K)}([0,1])$. Potom platí*

$$\arg \max_{x\in C^{(K)}([0,1])} KLP^{(K)}(x;P) = \theta$$

a zároveň

$$KLP^{(K)}(\theta;P) = \frac{\sum_{k=0}^{K} \alpha_k 2^{-(k+1)}}{\sum_{i=0}^{K} \alpha_i},$$

špeciálne pre vol'bu $\alpha = (1,\ldots,1)^T \in \mathbb{R}^{K+1}$

$$KLP^{(K)}(\theta;P) = \frac{1-2^{-(K+1)}}{K+1}.$$

Ďalej platí, že pre každé $x \in C^{(K)}([0,1])$ je $KLP^{(K)}((1-\alpha)\theta+\alpha x;P)$ nerastúca funkcia v argumente $\alpha \in [0,1]$.

Dôkaz. Podľa tvrdenia 4.2 o afinnej invariancii K-pásovej hĺbky môžeme bez ujmy na všeobecnosti predpokladať, rozdelenie P je angulárne symetrické okolo nulovej funkcie, pretože inak prejdeme k afinnej transformácii

$$T(x) = x - \theta$$

a rozdelenie je po takejto transformácii angulárne symetrické okolo nulovej funkcie (budeme značiť jednoducho 0), pričom hĺbka transformovaných funkcií sa nezmenila.

To, že mnohorozmerné rozdelenie je (angulárne) symetrické zrejme implikuje, že všetky jeho marginálne rozdelenia budú rovnako (angulárne) symetrické. Rozdelenie P je angulárne symetrické rozdelenie na $C^{(K)}([0,1])$. To znamená, že pre každé $t \in [0,1]$ a pre každé $k = 0,\ldots,K$ sú rozdelenia $P_t^{(0,\ldots,k)}$ na $\mathbb{R}^{k+1}$ angulárne symetrické okolo počiatku. Označme

$$g(\gamma) = \gamma x + (1-\gamma)\theta \in C^{(K)}([0,1]), \gamma \in [0,1].$$

Podľa tvrdenia 1.2 o maximalite a monotónii relatívnej voči najhlbšiemu bodu pre simplexovú hĺbku potom pre každú funkciu $x \in C^{(K)}([0,1])$ pre $0 \leq \gamma' \leq \gamma'' \leq 1$ za predpokladu, že θ je stredom angulárnej symetrie platí

$$\begin{aligned}
&\sum_{i=0}^{K} \alpha_i KLP^{(K)}\left(g\left(\gamma'\right);P\right) \\
&\qquad = \sum_{k=0}^{K} \alpha_k \int_0^1 SD\left(\gamma' x^{(0,\ldots,k)}(t) + \left(1-\gamma'\right)\theta^{(0,\ldots,k)}(t); P_t^{(0,\ldots,k)}\right) \mathrm{d}t \\
&\qquad \geq \sum_{k=0}^{K} \alpha_k \int_0^1 SD\left(\gamma'' x^{(0,\ldots,k)}(t) + \left(1-\gamma''\right)\theta^{(0,\ldots,k)}(t); P_t^{(0,\ldots,k)}\right) \mathrm{d}t \\
&\qquad = \sum_{i=0}^{K} \alpha_i KLP^{(K)}\left(g\left(\gamma''\right);P\right).
\end{aligned}$$

Tým sme dokázali časť tvrdenia o relatívnej monotónii a zároveň pre voľbu $\gamma' = 0$ a $\gamma'' = 1$ aj časť tvrdenia o maximalite v bode stredu symetrie. Ďalej opäť podľa tvrdenia 1.2 ak označíme $0_{k+1} = (0,\ldots,0)^T \in \mathbb{R}^{k+1}$ platí

$$\begin{aligned}
KLP^{(K)}(0;P) &= \frac{\sum_{k=0}^{K} KLP^k(0;P)\,\alpha_k}{\sum_{i=0}^{K}\alpha_i} \\
&= \frac{1}{\sum_{i=0}^{K}\alpha_i} \sum_{k=0}^{K} \alpha_k \int_0^1 SD\left(0_{k+1}; P_t^{(0,\ldots,k)}\right) \mathrm{d}t \\
&= \frac{1}{\sum_{i=0}^{K}\alpha_i} \sum_{k=0}^{K} \alpha_k \int_0^1 2^{-(k+1)} \mathrm{d}t \\
&= \frac{\sum_{k=0}^{K}\alpha_k 2^{-(k+1)}}{\sum_{i=0}^{K}\alpha_i}.
\end{aligned}$$

Prípad pre jednotkový vektor váh α je zrejmým dôsledkom predchádzajúceho. □

Ďalšou vlastnosťou, ktorú požadujeme od štatistickej hĺbkovej funkcie, je podmienka nulovosti v limite pre také postupnosti funkcií, ktorých norma rastie nad všetky medze. Táto podmienka však pre žiadnu integrálnu funkcionálnu hĺbku, teda hĺbku založenú na integrovaní konečnorozmerných hĺbok cez definičný obor funkcií podľa nejakej konečnej miery μ, nemôže byť splnená pre všetky spojite diferencovateľné funkcie. Existuje jednoduchý protipríklad postupnosti funkcií $\{g_n\}_{n=1}^{\infty} \subset C([-1,1])$ takých, že síce platí $\|g_n\|^{(K)} \xrightarrow[n\to\infty]{} \infty$, ale napriek tomu pre každú integrálnu hĺbku GS platí $GS(g_n) \xrightarrow[n\to\infty]{} \mu([-1,1])$.

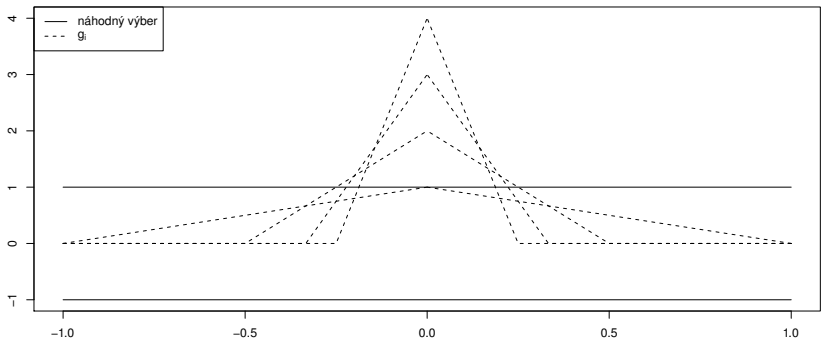
Príklad 14. Nech náhodný výber funkcionálnych dát je tvorený $n = 2$ konštantnými funkciami

$$x_1(t) = 1, \qquad x_2(t) = -1, \qquad t \in [-1,1].$$

Počítajme teraz postupne nejakú integrálnu funkcionálnu hĺbku tvorenú konečnou mierou μ na intervale $[-1,1]$ takou, že $\mu(0) = 0$ pre funkcie z postupnosti

$$g_n(t) = \begin{cases} 0 & \text{pre } |t| \geq 1/n, \\ -n^2 t + n & \text{pre } 0 \leq t \leq 1/n, \quad n \in \mathbb{N} \\ n^2 t + n & \text{pre } -1/n \leq t \leq 0. \end{cases}$$

Funkcie $g_1, \ldots, g_4$ sú znázornené na obrázku 4.1. Jedná sa o postupnosť spojitých



Obr. 4.1: Funkcie g_i pre $i = 1, \ldots, 4$.

lineárne lomených funkcií takých, že pre

$$t \notin A_n = [-1/n, 1/n]$$

platí $x_1(t) \leq g_n(t) \leq x_2(t)$ a zároveň $\|g_n\|^{(K)} = n \xrightarrow[n\to\infty]{} \infty$. Pretože ale pre každé $n \in \mathbb{N}$ platí $A_n \supset A_{n+1}$ a

$$\lim_{n\to\infty} \mu(A_n) = \mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \mu(0) = 0,$$

máme $GS(g_n) \xrightarrow[n\to\infty]{} \mu([-1,1])$, čo je v rozpore s predpokladom nulovosti hĺbky v limite. V prípade miery μ takej, že $\mu(0) \neq 0$ stačí uvažovať nejaký bod $t_0 \in [-1,1]$ taký, že $\mu(t_0) = 0$ a namiesto pôvodnej postupnosti uvažovať funkcie $\{g_n(t - t_0)\}_{n=1}^{\infty}$. Preto platí, že pre žiadnu integrálnu funkcionálnu hĺbku (K-pásovú hĺbku, Fraimanovu-Munizovej hĺbku alebo zovšeobecnenú pásovú hĺbku) neplatí v plnej všeobecnosti predpoklad nulovosti limity. Tým, že počítame mieru takých bodov definičného oboru funkcií, pre ktoré funkcie g_n ležia v páse tvorenom funkciami z náhodného výberu, integrálna hĺbka stráca kontrolu nad priebehom funkcií g_n na množinách, ktorých miera konverguje k nule.

Preto je možné utvoriť postupnosť funkcií g_n tak, aby v suprémovej norme konvergovali do nekonečna a aby body, v ktorých nadobúdajú suprémum boli práve v množinách, ktorých miera konverguje k nule. Analogický protipríklad bude zrejme možné

zostrojiť aj pre prípad funkcií z $C^{(K)}([0,1])$ pre ľubovoľné $K \in \mathbb{N}$ tak, aby pre ľubovoľné $k = 0,\ldots,K$ platilo $\sup_{t\in[0,1]} \left|g_n^{(k)}(t)\right| \xrightarrow[n\to\infty]{} \infty$ a zároveň aby $\sup_{t\in[0,1]} \left|g_n^{(l)}(t)\right|$ boli obmedzené pre každé $l \neq k$.

Predpoklad nulovosti limity teda zrejme nebude možné pre integrálne funkcionálne hĺbky splniť v prípade ak pripustíme, že funkcia, ktorej hĺbku hľadáme, môže v niektorej derivácií neobmedzene rásť.

Ukázali sme, že K-pásová hĺbka splňuje funkcionálnu analógiu vlastností P1 (tvrdenie 4.2), P2 a P3 (tvrdenie 4.5) z definície štatistickej hĺbkovej funkcie. Nesplňuje však vlastnosť P4 (príklad 14), rovnako ako žiadna integrálna funkcionálna hĺbka. Vlastnosť nulovosti v limite by geometricko-funkcionálna hĺbka splňovala, ak by sme postupovali v zmysle rozširovania pásovej hĺbky namiesto zovšeobecnenej pásovej hĺbky. Vo výberovom prípade by sme v 4.4 namiesto Lebesgueovej miery takých bodov $t \in [0,1]$, pre ktoré platí

$$x^{(0,\ldots,k)}(t) \in \mathbb{S}_{X_{i_1},\ldots,X_{i_{k+2}}}(t)$$

zaviedli tak ako pri pásovej hĺbke v 3.22 za mieru náležania vektorovej funkcie $x^{(0,\ldots,k)}$ do $(k+1)$-rozmernej množiny

$$\left\{\mathbb{S}_{X_{i_1},\ldots,X_{i_{k+2}}}(t), t \in [0,1]\right\}$$

jednoduchý indikátor.

Pozrime sa teraz na ďalšiu dôležitú vlastnosť, konzistenciu výberovej K-pásovej hĺbky ku svojmu populačnému náprotivku.

Tvrdenie 4.6. *Nech $K \in \mathbb{N}_0$. Označme ako $E \subset C^{(K)}([0,1])$ triedu takých funkcií $x \in C^{(K)}([0,1])$, pre ktoré existuje $M > 0$ také, že platí $\sup_{x\in E} \|x\|^{(K)} \leq M$ a zároveň $\left\{x^{(K)} : x \in E\right\}$ je rovnako spojitá trieda funkcií. Nech $x \in E$ a P je absolútne spojité rozdelenie pravdepodobnosti na E. Potom platí*

$$\sup_{x\in E} \left|KLP_n^{(K)}(x) - KLP^{(K)}(x)\right| \xrightarrow[n\to\infty]{s.i.} 0.$$

Dôkaz. Pretože sme predpokladali rovnakú obmedzenosť všetkých derivácií funkcií z E, máme

$$\sup_{x\in E} \left|KLP_n^{(K)}(x) - KLP^{(K)}(x)\right| = \sup_{x\in E, \|x\|^{(K)}\leq M} \left|KLP_n^{(K)}(x) - KLP^{(K)}(x)\right|$$

Na všetkých $(K+1)$ derivácií funkcie $x \in E$ sa môžeme pozerať aj naraz ako na vektorovú funkciu

$$x^{(0,\ldots,K)} : [0,1] \to \mathbb{R}^{K+1},$$

ktorej k-ta zložka bude predstavovať k-tu deriváciu funkcie x. Na takúto triedu vektorových funkcií

$$E^{(0,\ldots,K)} = \left\{x^{(0,\ldots,K)} : x \in E\right\}$$

použijeme Arzélovu-Ascoliho vetu, ktorá zaručí relatívnu kompaktnosť, a teda aj totálnu obmedzenosť triedy $E^{(0,\ldots,K)}$. Preto pre každé $\varepsilon > 0$ bude existovať konečná ε-sieť funkcií

$$\{x_1,\ldots,x_{N_\varepsilon}\} \subset C^{(K)}([0,1]) \tag{4.9}$$

taká, že bude platiť

$$E^{(0,\ldots,K)} \subset \bigcup_{i=1}^{N_\varepsilon} B(x_i, \varepsilon),$$

kde $B(x,\varepsilon)$ je otvorená guľa v priestore $C^{(K)}([0,1])$ o polomere ε a strede x. Podľa tvrdenia 4.4 ale v našom prípade absolútne spojitých marginálnych rozdelení P je K-pásová hĺbka $KLP^{(K)}$ spojitá na $C^{(K)}([0,1])$ a preto k dôkazu tvrdenia bude stačiť ukázať

$$\max_{i=1,\ldots,N_\varepsilon} \left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \xrightarrow[n\to\infty]{\text{s.i.}} 0. \tag{4.10}$$

K tomu, aby trieda vektorových funkcií $E^{(0,\ldots,K)}$ splňovala podmienky Arzélovej-Ascoliho vety, je potrebná rovnaká obmedzenosť a rovnaká spojitosť všetkých zložiek vektorových funkcií. Označme teda triedu funkcií tvoriacich k-tu zložku triedy $E^{(0,\ldots,K)}$ ako

$$E^k = \left\{ x^{(k)} : x \in E \right\}$$

pre $k = 0,\ldots,K$. Rovnakú obmedzenosť triedy E^k sme predpokladali, stačí teda dokázať rovnakú spojitosť. Táto však pre triedu funkcií E^k, $k = 0,\ldots,K-1$ plynie z rovnakej obmedzenosti triedy E^{k+1}, pretože každá funkcia triedy E^k má deriváciu obmedzenú konštantou M. Podľa vety o strednej hodnote teda platí že všetky funkcie triedy E^k sú rovnako globálne lipschitzovské, teda aj rovnako spojité. Pre triedu funkcií E^K sme rovnakú spojitosť museli predpokladať. Tým sme ukázali, že pre každú triedu E^k sú splnené predpoklady Arzélovej-Ascoliho vety, a preto trieda funkcií $E^{(0,\ldots,K)}$ je relatívne kompaktná v priestore $C^{(K)}([0,1])$.

Pre každú konečnú ε-sieť funkcií 4.9 teraz stačí dokázať platnosť 4.10. Platí

$$\begin{aligned}
&P\left(\max_{i=1,\ldots,N_\varepsilon} \left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \geq \varepsilon \right) \\
&\qquad = P\left(\bigcup_{i=1}^{N_\varepsilon} \left[\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \geq \varepsilon \right] \right) \\
&\qquad \leq \sum_{i=1}^{N_\varepsilon} P\left(\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \geq \varepsilon \right) \\
&\qquad \leq N_\varepsilon \max_{i=1,\ldots,N_\varepsilon} P\left(\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \geq \varepsilon \right) \\
&\qquad \leq N_\varepsilon \max_{i=1,\ldots,N_\varepsilon} \frac{\mathrm{E}\left[\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right|^4 \right]}{\varepsilon^4},
\end{aligned} \tag{4.11}$$

kde posledná nerovnosť platí podľa Čejbyševovej nerovnosti. Podľa tvrdenia 4.1 je ale $KLP_n^{(K)}$ U-štatistika a preto s využitím odhadu pre štvrtý centrálny moment U-štatistiky (napr. Serfling [23, Lemma 5.2.2A]) môžeme nakoniec 4.11 odhadnúť

$$\begin{aligned}
&N_\varepsilon \max_{i=1,\ldots,N_\varepsilon} \frac{\mathrm{E}\left[\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right|^4 \right]}{\varepsilon^4} \\
&\qquad = \frac{N_\varepsilon}{\varepsilon^4} \max_{i=1,\ldots,N_\varepsilon} \mathrm{E}\left[\left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right|^4 \right] = O\left(n^{-2}\right).
\end{aligned}$$

Ak pre každé $\varepsilon > 0$ definujme postupnosť náhodných javov

$$A_n = \left\{ \omega \in \Omega : \max_{i=1,\ldots,N_\varepsilon} \left| KLP_n^{(K)}(x_i) - KLP^{(K)}(x_i) \right| \geq \varepsilon \right\},$$

platí podľa predchádzajúceho

$$\sum_{n=1}^{\infty} P(A_n) \leq \sum_{n=1}^{\infty} O\left(n^{-2}\right) < \infty.$$

Použitím Cantelliho lemmy (pozri napríklad Dupač a Hušková [6, Věta 4.1]) dostávame nakoniec

$$P\left(\limsup_{n\to\infty} A_n \right) = 0,$$

čo už dáva silnú konzistenciu $KLP_n^{(K)}$, takže

$$\sup_{x\in E} \left| KLP_n^{(K)}(x) - KLP^{(K)}(x) \right| \xrightarrow[n\to\infty]{\text{s.i.}} 0.$$

□

Tvrdenie 3.1 spoločne s tvrdením 4.6 nám zaručuje (za predpokladu splnenia predpokladov tvrdenia 4.6) konvergenciu výberového useknutého priemeru počítaného na základe K-pásovej hĺbky k svojmu populačnému náprotivku.

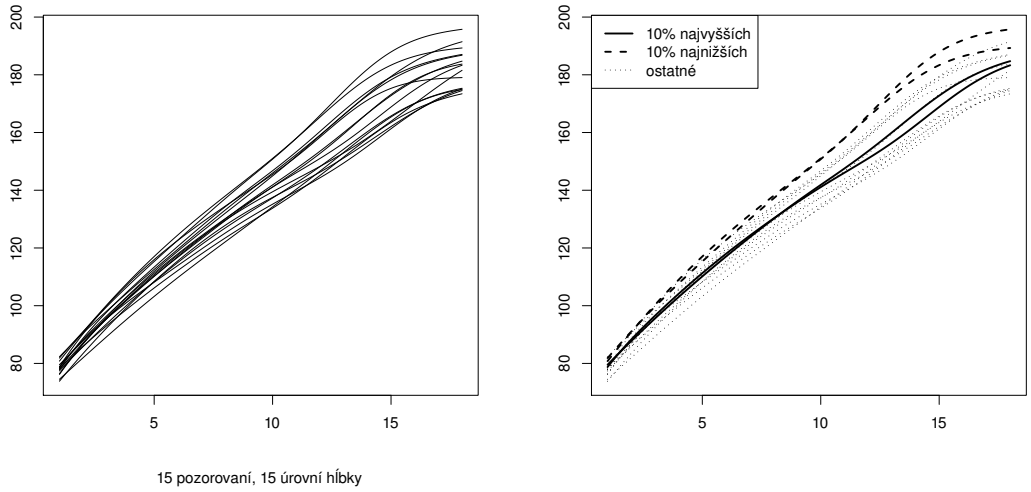
Aplikujme na záver 1-pásovú hĺbku ($K = 1$) s voľbou koeficientov $\alpha = (1,1)^T$ na náhodné výbery zo všetkých predchádzajúcich príkladov funkcionálnych dát.

Príklad 15. Tak ako v príkladoch 9 a 12 preveďme hĺbkovú analýzu dát tvorených rastovými krivkami chlapcov pomocou 1-pásovej hĺbky, ktorá na rozdiel od geometrických hĺbok rozoznáva aj pozorovania odľahlé v tvare. Ako vidíme na obrázku 4.2, výsledky sú vizuálne úplne totožné s tými na obrázku 3.4, teda je zrejmé, že rozšírenie inferencie na prvú deriváciu nemení v tomto prípade výsledky analýzy. To je ekvivalentné tomu, že v náhodnom výbere sa nenachádzajú funkcie netypické v tvare.

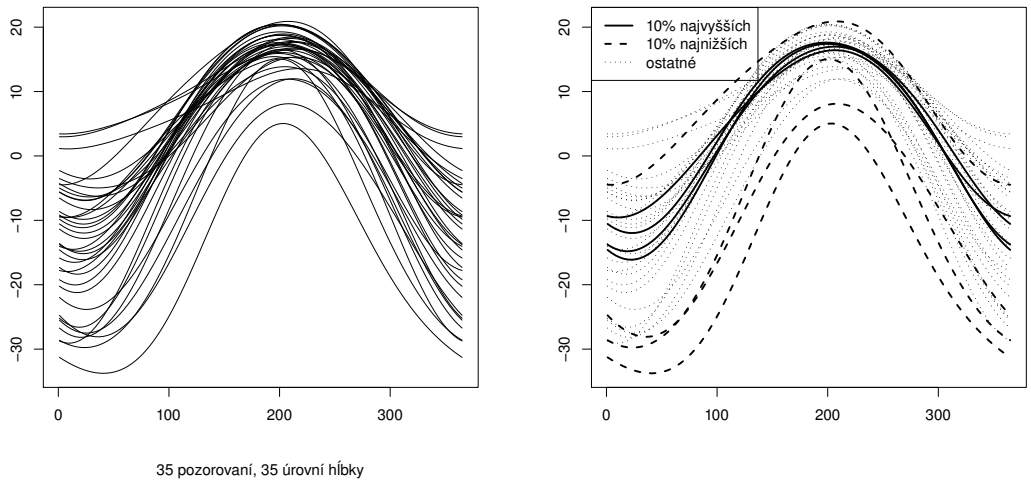
Príklad 16. Ak analyzujeme náhodný výber dát s priemernými teplotami v 35 kanadských mestách (príklady 5 a 8) pomocou 1-pásovej hĺbky (obrázok 4.3), dostávame na rozdiel od príkladu 15 už aj vizuálne odlišné výsledky ako pri analýze pomocou ostatných hĺbkových funkcií. Zmena oproti indukovanej polopriestorovej hĺbke je viditeľná najmä v počte úrovní hĺbky, pretože pri použití indukovanej hĺbky dostávame iba 4 úrovne, zatiaľ čo pri použití 1-pásovej hĺbky dostávame rovnako ako v príklade 8 dokonalé rozlíšenie funkcií, 35 úrovní hĺbky pre 35 pozorovaní.

Ďalší veľký rozdiel je v tom, že zatiaľ čo pri použití indukovanej polopriestorovej hĺbky dostali takmer všetky funkcie nulovú hĺbku a boli teda označené za odľahlé, v prípade Fraimanovej-Munizovej hĺbky a rovnako aj pri použití 1-pásovej hĺbky je kandidát na odľahlú funkciu, teda pozorovanie s nulovou hĺbkou, iba jediné. Jedná sa samozrejme o funkciu, ktorá celým svojím grafom leží pod grafmi všetkých ostatných funkcií.

Ak ďalej porovnávame Fraimanovu-Munizovej hĺbku s 1-pásovou hĺbkou, vidíme iba malé rozdiely pri funkciách s najväčšou hĺbkou. Pri použití 1-pásovej hĺbky získavajú väčšiu hodnotu pozorovania podobné náhodnému výberu aj v raste.



Obr. 4.2: 1-pásová hĺbka a rast chlapcov.



Obr. 4.3: 1-pásová hĺbka a kanadské počasie.

Príklad 17. Zatiaľ čo v príklade 15 sme nevideli žiaden zrejmý rozdiel medzi geometrickými hĺbkami a K-pásovou hĺbkou a v príklade 16 sme rozdiely už objavili, ale stále sme nemohli jednoznačne porovnať vhodnosť prístupov, na simulovaných dátach s netypickou funkciou (pozri príklady 6, 10 a 13) uvidíme veľké rozdiely vo výsledkoch hĺbkových analýz. Hlavnou nevýhodou konceptu indukovanej hĺbky je nízke rozlíšenie pozorovaní na rôzne úrovne hĺbky (v príklade 6 iba 2 z 11). Na druhej strane však funkcionálne hĺbky (aspoň za istých predpokladov) boli schopné rozlíšiť funkciu odľahlú v tvare, zatiaľ čo žiadna z geometrických hĺbkových funkcií toto pozorovanie neodlíšila.

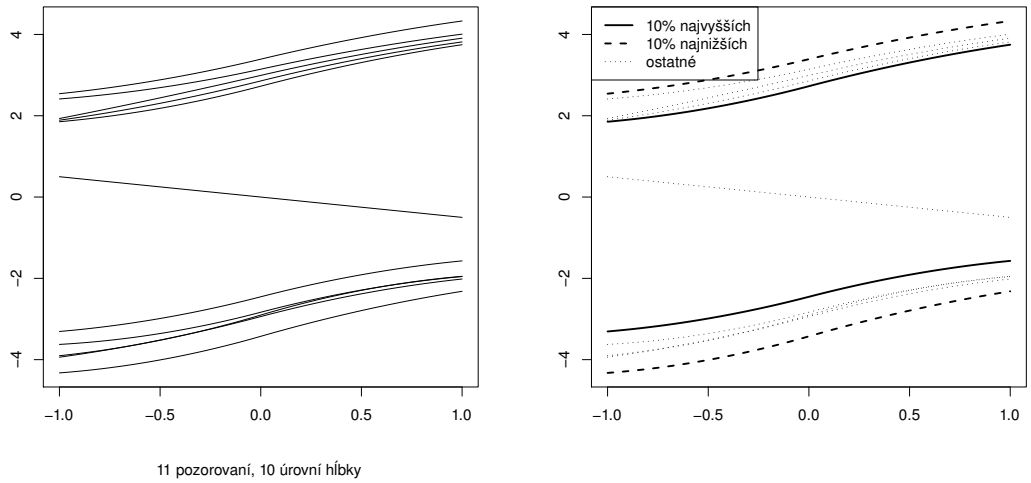
Ak však na náhodný výber simulovaných dát použijeme tzv. *priemernú* 1*-pásovú hĺbku*, teda 1-pásovú hĺbku s váhami $\alpha = (1, 1)^T$ (obrázok 4.4), vidíme, že oba tieto nedostatky 1-pásová hĺbka odstránila tak, ako sme to očakávali. To, že funkcie nie sú dokonale rozlíšené (10 úrovní z 11) je spôsobené iba tým, že obe funkcie s nulovou hĺbkou ležia celým svojím grafom mimo pásu tvoreného zvyšnými funkciami už v nul-

tej derivácií. Preto museli obe dostať zhodne nulovú hĺbku. Vidíme, že K-pásová hĺbka rozlišuje pozorovania dokonale rovnako ako geometrické hĺbky.

Na druhej strane ako pozorovania s najväčšou hodnotou hĺbky sú označené dve pozorovania ležiace najbližšie (v rámci metriky 4.1) k funkcii aritmetického priemeru a pritom splňujú podmienku, že sa nelíšia v tvare od ostatných funkcií. Naproti tomu netypická funkcia, ktorá bola v prípade geometrických hĺbok vždy označená za funkcionálnu analógiu mediánu, už nie je medzi 10 % funkcií s najväčšou hĺbkou. Napriek tomu, že je jasne odľahlá v tvare, nie je pri analýze 1-pásovou hĺbkou označená za kandidáta na odľahlé pozorovanie. To je spôsobené tým, že je to zároveň najtypickejšia funkcia náhodného výberu v polohe. Nultý sčítanec 1-pásovej hĺbky teda dosahuje najvyššej hodnoty.

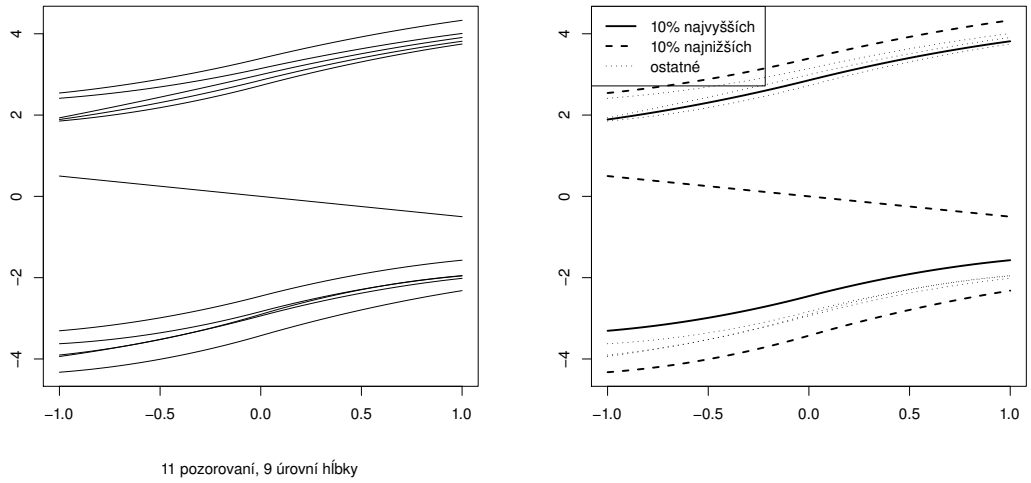
Pre inú voľbu parametrov $\alpha = (0,1)^T$ (tzv. *obmedzená* 1-*pásová hĺbka*) vidíme vo výsledku (obrázok 4.5), že všetky tri netypické funkcie: dve tvoriace obal náhodného výberu a klesajúca funkcia, dostávajú najnižšiu možnú, nulovú hĺbku. Takáto voľba parametrov zodpovedá tomu, že je kladený dôraz na odhalenie pozorovaní odľahlých ako v polohe tak aj v tvare, pričom odhalenie odľahlosti pozorovaní v polohe a odľahlosti pozorovaní v tvare má rovnakú prioritu. Naproti tomu v prípade priemernej 1-pásovej hĺbky ($\alpha = (1,1)^T$) pozorovania odľahlé v polohe sú považované za „extrémnejšie“ ako pozorovania odľahlé v tvare ale nie v polohe.

Ukázali sme, že nevýhody ako funkcionálnych (zlá rozlíšiteľnosť), tak aj geometrických (identifikácia funkcií odľahlých v tvare) hĺbok je možné odstrániť použitím vhodne volenej geometricko-funkcionálnej K-pásovej hĺbky.



Obr. 4.4: Priemerná 1-pásová hĺbka a simulované dáta.

V diskusii k príkladu 17 sme naznačili, že napriek tomu, že K-pásová hĺbka odlišuje funkcie odľahlé v tvare, v prípade voľby váh $\alpha = (1,\ldots,1)^T$ dostávajú takéto funkcie (odľahlé napríklad iba v l-tej derivácií pre $l \in \mathbb{N}$) stále pomerne vysokú hodnotu K-pásovej hĺbky. To je spôsobené tým, že nultý až $(l-1)$-ty sčítanec stále neindikujú odľahlosť. Preto aj celková hĺbka môže byť vysoká. Túto disproporciu je možné odstrániť jednoduchou voľbou vektoru váh $\alpha = (0,\ldots,0,1)^T$ tak, ako v príklade 17. Tá zabezpečí, že funkcie odľahlé v niektorej z derivácií rádov $0,\ldots,K$ budú odhalené a identifikované ako odľahlé s rovnakou prioritou.



Obr. 4.5: Obmedzená 1-pásová hĺbka a simulované dáta.

Pomocou rôznych volieb vektoru váh môžeme tiež zisťovať, v ktorej derivácií je ukrytá aká podstatná časť K-pásovej hĺbky. Napríklad pre funkciu s veľmi nízkou hĺbkou teda môžeme usúdiť, pre poruchu v ktorej derivácií je funkcia identifikovaná ako možné odľahlé pozorovanie.

Kapitola 5

Klasifikácia funkcionálnych dát

Vyšetrime teraz na jednoduchej simulačnej štúdii možnosti praktického použitia zavedených hĺbkových funkcionálov. Budeme riešiť niekoľko úloh riadenej klasifikácie simulovaných aj reálnych funkcionálnych dát. Popíšme preto na úvod, v čom problém riadenej klasifikácie funkcionálnych dát spočíva.

V plnej všeobecnosti predpokladajme, že máme dané známe *tréningové skupiny* funkcií $G_1, G_2, \ldots, G_M$, kde i-ta tréningová skupina ($i = 1, \ldots, M$) je náhodný výber o rozsahu $n_i \in \mathbb{N}$ pochádzajúci z rozdelenia pravdepodobnosti $P_i \in \mathcal{P}\left(C^{(K)}([0,1])\right)$, $P_i \neq P_j$ pre $i \neq j$. Ďalej nech $X \in C^{(K)}([0,1])$ je náhodná funkcia nezávislá na všetkých tréningových skupinách, pochádzajúca z niektorého rozdelenia pravdepodobnosti P_m, pričom m je neznáme. Našou úlohou je rozhodnúť, z ktorého rozdelenia pravdepodobnosti nové pozorovanie pochádza. Na to sa môžeme pozerať aj ako na hľadanie M disjunktných množín $A_1, A_2, \ldots, A_M \subset C^{(K)}([0,1])$ (nie nutne $C^{(K)}([0,1]) = \bigcup_{i=1}^{M} A_i$) takých, že ak $X \in A_i$, potom funkcii X priradíme distribúciu P_i.

Čím lepšie je klasifikačné pravidlo, tým menej náhodných funkcií je priradených k nesprávnemu modelu. Dobré klasifikačné pravidlo by tiež malo udržovať úroveň neklasifikovaných funkcií čo najbližšie k nule.

Niekoľko klasifikačných pravidiel známych z konečnorozmerných modelov bolo rozšírených na prípad funkcionálnych dát. Napríklad známe pravidlo najbližších susedov (pozri Devroye et al. [5] alebo Stone [26] pre konečnorozmerný a Biau et al. [1] alebo Cérou a Guyander [2] pre funkcionálny prípad) Cuevas et al. [4] porovnáva s ďalšími klasifikačnými pravidlami funkcionálnych dát. Jedná sa o zovšeobecnenie pravidla vzdialenosti funkcie od priemeru tréningovej skupiny.

Väčšinu klasifikačných pravidiel je možné jednoduchým spôsobom robustizovať vnesením informácie o hĺbke funkcií voči tréningovým skupinám ako píšu López-Pintado a Romo [14]. Popíšme si teraz základnú myšlienku tejto techniky na niekoľkých príkladoch.

Nech $d(.,.)$ je nejaká metrika na priestore $C^{(K)}([0,1])$ (typicky metrika indukovaná normou 4.1). Funkciu X priradíme k rozdeleniu pravdepodobnosti P_m podľa *pravidla vzdialenosti od useknutého priemeru*, ak platí

$$m = \arg \min_{i=1,\ldots,M} d(X; \widehat{\overline{X}^{i}_{n_i\beta_i}}),$$

kde $\widehat{\overline{X}^{i}_{n_i\beta_i}}$ označuje β_i-useknutý priemer (pozri 3.3) tréningovej skupiny G_i voči nejakej pevnej hĺbke D a konštante β_i volenej tak, aby bolo do výpočtu useknutého priemeru

v *i*-tej tréningovej skupine zahrnutých $n_i - \lfloor n_i\alpha \rfloor$ pozorovaní pre pevné $\alpha \in (0,1)$. Takéto robustizovanie je vhodné najmä v prípade, ak máme podozrenie z kontaminácie tréningových skupín malým počtom odľahlých funkcií. V jednoduchom prípade nekontaminovaných dát však tento prístup zrejme zbytočne zvyšuje variabilitu nesprávneho priradenia voči pravidlu vzdialenosti od priemeru.

López-Pintado a Romo [14] ďalej popísujú *pravidlo vzdialenosti od náhodného výberu* spočívajúce v nájdení takej tréningovej skupiny $G_m = \{x_{m1}, \ldots, x_{mn_m}\}$, pre ktorú je minimalizovaný súčet

$$AD(X, G_m) = \frac{\sum_{i=1}^{n_m} d(X, x_{mi})}{n_m}. \tag{5.1}$$

Ak nahradíme aritmetický priemer v 5.1 priemerom váženým hĺbkou pozorovaní voči náhodnému výberu G_m, dostávame robustizovanú verziu kritéria 5.1. Preto funkcie náhodného výberu, ktoré sú podozrivé z toho byť odľahlé, dostanú malé váhy a hodnota funkcionálu *AD* bude iba málo ovplyvnená takýmito pozorovaniami. Rovnako ako prvé rozhodovacie pravidlo založené na hĺbke, aj pravidlo váženej vzdialenosti od náhodného výberu je výhodné používať v prípade kontaminovaných tréningových skupín. Pri použití na nekontaminované modely sa váženie v 5.1 stane nevýznamné a výsledky budú podobné ako pri použití pôvodného pravidla vzdialenosti od náhodného výberu.

Pre náš účel vzájomného porovnania funkcionálnych hĺbok predstavených v predchádzajúcich kapitolách najlepšie poslúži najjednoduchšie *pravidlo maximálnej hĺbky*: pri danej výberovej funkcionálnej hĺbke D_n funkciu X priradíme k distribúcii P_m práve vtedy keď

$$m = \arg \max_{i=1,\ldots,M} D_{n_i}(X; G_i).$$

V prípade, že funkcia X dostane voči niekoľkým rôznym tréningovým skupinám rovnakú, najvyššiu hodnotu hĺbky, funkciu neklasifikujeme. Alternatívne je však možné klasifikáciu randomizovať tým, že by sme sa medzi skupinami s najväčšou hodnotou hĺbky rozhodli náhodne. Pretože ale v našich simuláciách budú vždy iba dve skupiny, nebudeme randomizáciu používať.

Dôvodom prečo práve pravidlo hodnoty hĺbky slúži pre účely porovnania hĺbkových funkcií najlepšie je, že dobrá hĺbková funkcia by mala byť schopná rozlíšiť rozdielne vlastnosti uvažovaných distribúcií. Ak má funkcia X podobné funkcionálne vlastnosti ako funkcie z niektorej tréningovej skupiny, hĺbka by to mala rozoznať a priradiť X vysokú hodnotu vzhľadom k podobnej tréningovej skupine. Aby sme zabezpečili porovnateľnosť hĺbok voči rôznym tréningovým skupinám, v každom prípade bude rozsah každej tréningovej skupiny rovný konštante $n \in \mathbb{N}$.

Vo všetkých nasledovných simuláciách porovnáme 5 hĺbkových funkcionálov:

- *pásové hĺbky* pre odporúčané hodnoty $J = 2$ ($LP_n^{(2)}$) a $J = 3$ ($LP_n^{(3)}$) (aproximované pomocou resamplingovej metódy do skupín o veľkosti 10 ako navrhli López-Pintado a Jornsten [13], pozri aj kapitolu 6),

- *Fraimanovu-Munizovej hĺbku* ekvivalentnú *zovšeobecnenej pásovej hĺbke* pre $J = 2$ a tiež $J = 3$ ($GLP_n^{(2)}$)

- 1-pásovú hĺbku s dvomi alternatívami voľby váh: *priemernú* 1-*pásovú hĺbku* pre voľbu $\alpha = (1,1)^T$ ($A1KLP_n$) a *obmedzenú* 1-*pásovú hĺbku* pre voľbu $\alpha = (0,1)^T$ ($R1KLP_n$).

Poznamenajme, že v simuláciách nebudú použité indukované hĺbky, pretože ani v jednom prípade dáta nebudú vyhladzované do konenčnorozmerného priestrou.

Vo všetkých simuláciách porovnáme vyššie popísané hĺbkové funkcionály v úlohe riadenej klasifikácie s $M = 2$ rozdeleniami pravdepodobnosti na spojitých funkciách na kompaktnom intervale. Kľúčové indikátory pri porovnávaní hĺbok v klasifikačnej úlohe sú pravdepodobnosť nesprávnej klasifikácie, pravdepodobnosť toho, že funkcia nebude klasifikovaná, a ako doplnok pravdepodobnosť správnej klasifikácie. Pravdepodobnosť toho, že funkcia nebude klasifikovaná odhadneme pomerom počtu neklasifikovaných funkcií voči počtu všetkých funkcií, ktoré mali byť klasifikované. Ostatné pravdepodobnosti odhadneme analogicky.

Ďalšie technické detaily simulácií sú:

- *Počet opakovaní:* Každá simulačná štúdia je založená na 100 nezávislých opakovaniach.
- *Rozsah tréningových a testovaných skupín:* V každom opakovaní je z oboch distribúcií P_1 a P_2 generovaný náhodný výber rozsahu 100 pozorovaní. Podobne sú generované dve testované skupiny o rozsahu 50 pozorovaní z oboch distribúcií nezávisle na tréningových skupinách.
- *Diskretizácia:* Všetky funkcie sú prevedené do diskretizovanej formy v 51 ekvidistantných bodoch na intervale $[0,1]$. To znamená, že každá funkcia je reprezentovaná 51-rozmerným vektorom funkčných hodnôt na mriežke.
- *Derivácie funkcií:* Derivácie použité pre výpočet 1-pásových hĺbok sú získané vyhladzovaním metódou najmenších štvorcov pre kubické spline funkcie s 51 uzlami v bodoch mriežky. Vyhladená funkcia je následne derivovaná a výsledná derivácia je znovu diskretizovaná do bodov mriežky.
- *Interpretácia výsledkov:* Pre každú uvažovanú hĺbku ukazujú dve tabuľky hlavné popisné štatistiky distribúcie pomeru správne klasifikovaných a neklasifikovaných funkcií v 100 opakovaniach. Rovnaké výsledky sú pre názornosť vykreslené aj ako krabicové diagramy korešpondujúce s tabuľkami.

5.1 Porovnanie na základe simulácií - model posunutia v polohe

Čo by správna hĺbka určite mala postihnúť v prípade rozdelení pravdepodobnosti na funkcionálnych priestoroch je rozdielnosť v polohe a v tvare funkcií stredných hodnôt oboch distribúcií. Preto v našich simuláciách budeme uvažovať modely, ktoré sa líšia iba vo funkcii strednej hodnoty. V oboch distribúciách budú pozorovania generované podľa modelu

$$X(t) = m(t) + \varepsilon(t), t \in [0,1], \tag{5.2}$$

kde $m(t)$ je spojitá a aspoň raz diferencovateľná funkcia strednej hodnoty a $\varepsilon(t)$ je gaussovský proces s nulovou strednou hodnotou a vo všetkých prípadoch rovnakou autokovariančnou funkciou

$$\mathrm{Cov}(s,t) = 0.2 \exp\left(-\frac{|s-t|}{0.3}\right), s,t \in [0,1].$$

Oba porovnávané procesy majú teda rovnakú náhodnú štruktúru a jediný rozdiel medzi nimi sa dá nájsť v posunutí alebo tvare funkcií stredných hodnôt.

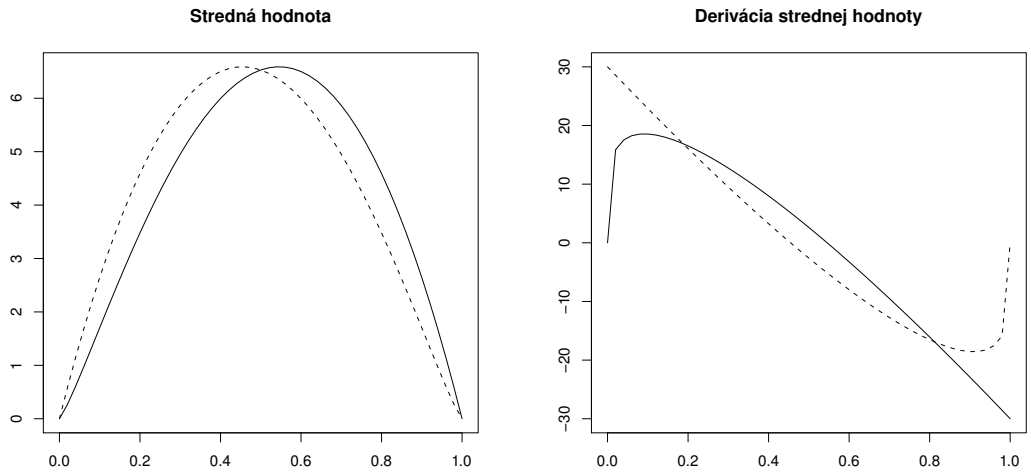
Prvý model, ktorý budeme uvažovať, bude identický prvému modelu použitému v štúdii Cuevas et al. [4]. Rozdiel v strednej hodnote oboch procesov sa dá popísať ako posunutie v polohe, nie však v tvare. Funkcia strednej hodnoty rozdelenia P_1 je

$$m_1(t) = 30(1-t)t^{1.2}$$

a stredná hodnota rozdelenia P_2 je podobného tvaru

$$m_2(t) = 30t(1-t)^{1.2}.$$

Obe funkcie spolu so svojimi prvými deriváciami sú vykreslené na obrázku 5.1.



Obr. 5.1: Funkcie strednej hodnoty a ich prvé derivácie v modele 1.

Vidíme, že sa líšia iba v nepatrnom posunutí v polohe a nezdá sa, že by sa ich prvé derivácie podstatne líšili na veľkej časti intervalu $[0,1]$. Preto sa zdá, že dodatočná informácia vnesená do výpočtu hĺbky cez aproximáciu prvej derivácie funkcií nebude výrazne meniť klasifikačné proporcie. Mohlo by sa dokonca zdať, že skutočným efektom pridania prvej derivácie do výpočtov budú o trochu horšie výsledky, pretože pridaním informácie o málo rozdielnych deriváciách iba zbytočne zvýšime variabilitu charakteristík klasifikácie.

V tabuľkách 5.1 a 5.2 vidíme základné popisné štatistiky odhadu pravdepodobnosti správnej klasifikácie (tabuľka 5.1) a odhadu pravdepodobnosti, že krivka nebude klasifikovaná (tabuľka 5.2). Príslušné krabicové diagramy pomeru správne klasifikovaných a neklasifikovaných kriviek nájdeme na obrázku 5.2.

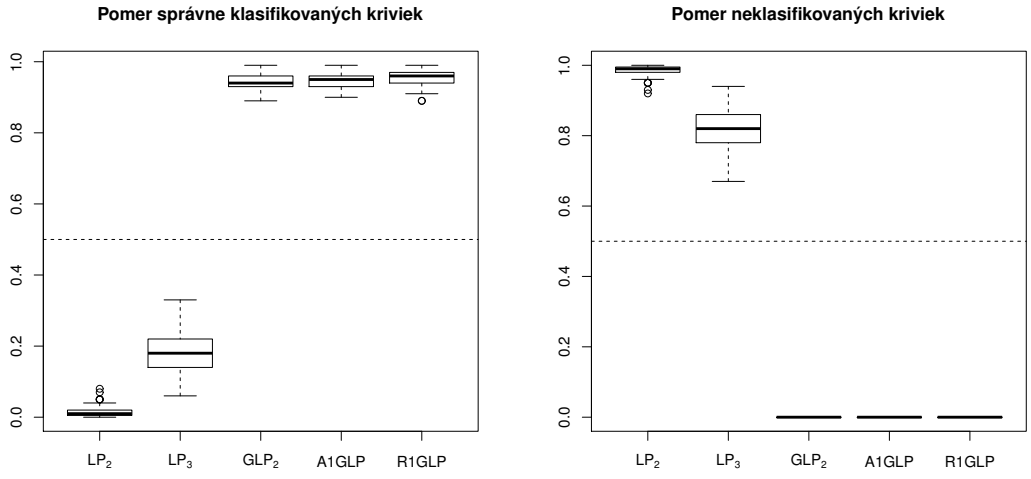
Hneď prvý zaujímavý jav ktorý vidíme je katastrofálne zlý pomer správne klasifikovaných kriviek pre obe pásové hĺbky $LP_n^{(2)}$ a tiež $LP_n^{(3)}$. To je spôsobené šumom vneseným do funkcionálnych pozorovaní. Gaussovský proces generujúci funkcie totiž spôsobuje, že pre každú funkciu je pravdepodobnosť javu, že na nejakom malom intervale jej graf vykĺzne z pásu tvoreného všetkými funkciami tréningovej skupiny, veľmi vysoká. To môžeme pozorovať aj na obrázku 5.3, kde vidíme 5 trajektórií náhodných funkcií a ich derivácií z rozdelenia P_1 a 5 z rozdelenia P_2. Preto každá neintegrálna hĺbka v tomto prípade prisudzuje funkcii triviálne nulovú hodnotu hĺbky

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0	0.06	0.89	0.9	0.89
1st Qu.	0.0075	0.14	0.93	0.93	0.94
Median	0.01	0.18	0.94	0.95	0.96
Mean	0.0164	0.1823	0.9404	0.947	0.956
3rd Qu.	0.02	0.22	0.96	0.96	0.97
Max.	0.08	0.33	0.99	0.99	0.99

Tabuľka 5.1: Pomer správne klasifikovaných kriviek v modele 1.

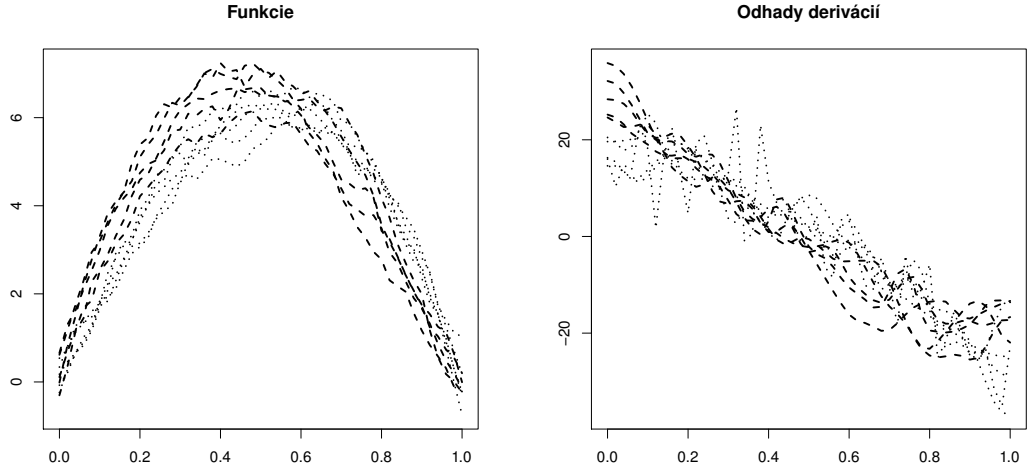
	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0.92	0.67	0	0	0
1st Qu.	0.98	0.78	0	0	0
Median	0.99	0.82	0	0	0
Mean	0.9836	0.817	0	0	0
3rd Qu.	0.9925	0.86	0	0	0
Max.	1	0.94	0	0	0

Tabuľka 5.2: Pomer neklasifikovaných kriviek v modele 1.



Obr. 5.2: Krabicové diagramy pre model 1.

a funkcia zostáva neklasifikovaná. Ako vidíme, napriek tomu, že takmer každá funkcia klasifikovaná neintegrálnou hĺbkou bola klasifikovaná správne, celkový pomer správne klasifikovaných funkcií je v podstate zanedbateľný.



Obr. 5.3: Niekoľko realizácií procesov v modele 1.

Na druhú stranu v prípade integrálnych hĺbkových funkcionálov je tento fenomén veľmi nepravdepodobný. V takomto prípade by posudzovaná funkcia X preto, aby jej bola určená nulová hodnota hĺbky, musela ležať celým svojím grafom mimo pásu tvoreného všetkými funkciami náhodného výberu.

Porovnajme teraz zovšeobecnenú pásovú hĺbku $GLP_n^{(2)}$ s dvomi verziami 1-pásovej hĺbky $A1KLP_n$ a $R1KLP_n$. Napriek tomu že sa na prvý pohľad nezdalo, že by prvé derivácie funkcií mali spôsobiť výrazný rozdiel v hĺbkach pre funkcie z oboch rozdelení pravdepodobnosti, pomer správne klasifikovaných funkcií je o niečo lepší pre pravidlá založené na 1-pásových hĺbkach. Preto aj keď zovšeobecnené pásové hĺbky dosahujú uspokojivé výsledky, môžeme vyhlásiť obmedzenú 1-pásovú hĺbku za najlepšie klasifikujúci hĺbkový funkcionál.

5.2 Porovnanie na základe simulácií - modely posunutia v tvare

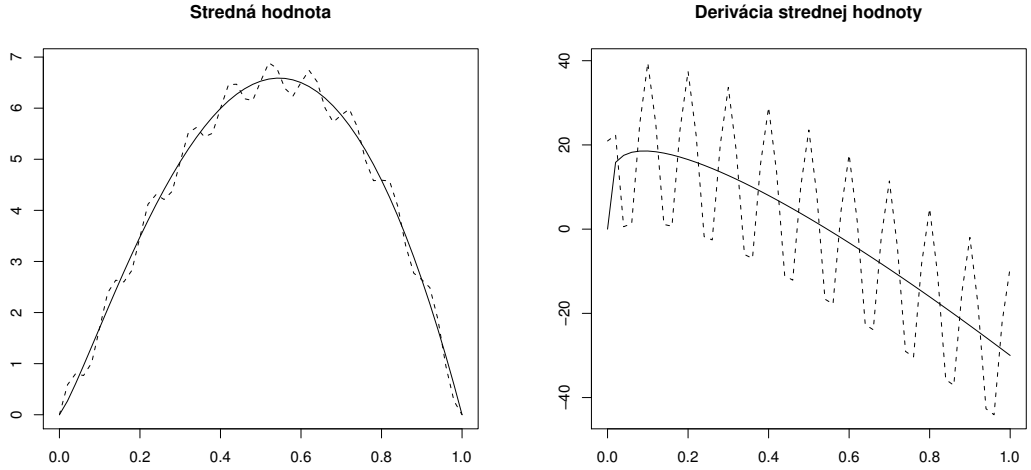
Podobne ako v modele 1, budeme aj v modele 2 porovnávať hĺbkové funkcionály pri úlohe klasifikácie medzi dvomi rozdielnymi rozdeleniami pravdepodobnosti tvorenými procesom tvaru 5.2. Jediný rozdiel oproti predchádzajúcemu bude v tom, že funkcie strednej hodnoty sa nebudú veľmi líšiť v polohe, ale najmä v tvare. To znamená, že napriek tomu, že si trajektórie procesov z oboch rozdelení budú veľmi blízko v suprémovej metrike, ich derivácie už budú vykazovať podstatnú rozdielnosť. Stredná hodnota rozdelenia pravdepodobnosti P_1 zostane v tvare

$$m_1(t) = 30(1-t)t^{1.2} \tag{5.3}$$

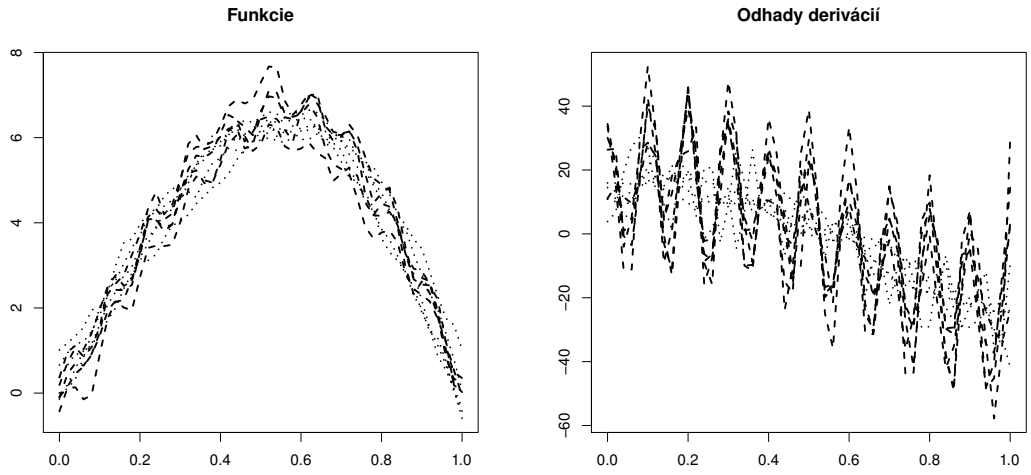
ako v modele 1. Funkcia strednej hodnoty P_2 bude oscilujúca verzia 5.3

$$m_2(t) = 30(1-t)t^{1.2} + \frac{\sin(20\pi t)}{3}.$$

Funkcie oboch stredných hodnôt a ich derivácie môžeme vidieť na obrázku 5.4. Niekoľko realizácií oboch procesov vidíme na obrázku 5.5.



Obr. 5.4: Funkcie strednej hodnoty a ich prvé derivácie v modele 2.



Obr. 5.5: Niekoľko realizácií procesov v modele 2.

Ak teraz vykonáme identickú simuláciu ako v modele 1 s inými funkciami strednej hodnoty, dostávame výsledky v zobrazené v tabuľke 5.3 (pomer správne klasifikovaných kriviek), tabuľke 5.4 (pomer neklasifikovaných kriviek) a na obrázku 5.6 (krabicové diagramy).

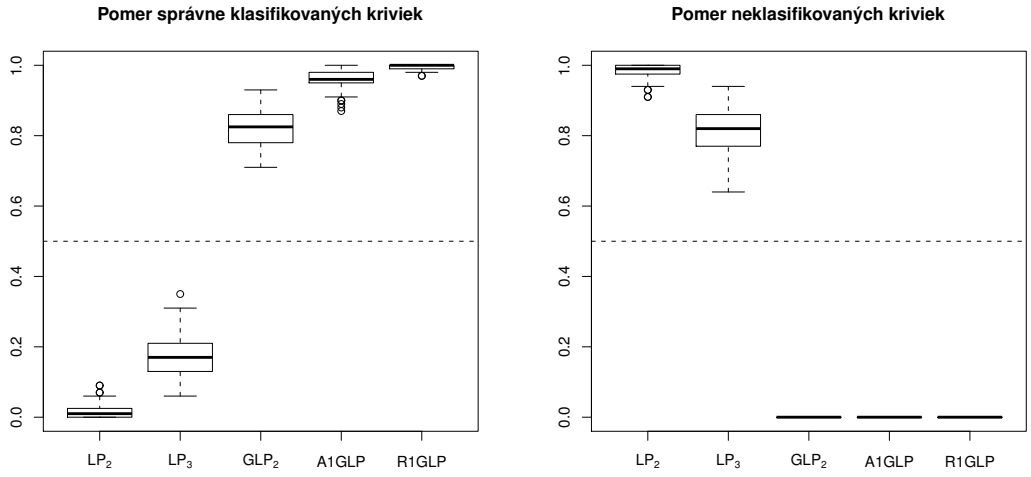
Výsledky pre obe verzie pásovej hĺbky sú veľmi podobné tým z modelu 1 (veľmi malá pravdepodobnosť správnej klasifikácie). Na rozdiel od simulácií v modele 1 však vidíme, že zovšeobecnená pásová hĺbka sa zdá byť oveľa horšou voľbou ako v modele 1. To je spôsobené tým, že zovšeobecnené pásové hĺbky zrejme nie sú schopné identifikovať funkcie odľahlé v tvare ako sme už naznačili v diskusii v kapitole 3. Ak sú pozorovania z oboch distribúcií blízko v suprémovej metrike, ale jedna z dis-

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0	0.06	0.71	0.87	0.97
1st Qu.	0	0.13	0.78	0.95	0.99
Median	0.01	0.17	0.825	0.96	1
Mean	0.0163	0.1724	0.821	0.9609	0.9935
3rd Qu.	0.0225	0.21	0.86	0.98	1
Max.	0.09	0.35	0.93	1	1

Tabuľka 5.3: Pomer správne klasifikovaných kriviek v modele 2.

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0.91	0.64	0	0	0
1st Qu.	0.9775	0.77	0	0	0
Median	0.99	0.82	0	0	0
Mean	0.9832	0.8193	0	0	0
3rd Qu.	1	0.86	0	0	0
Max.	1	0.94	0	0	0

Tabuľka 5.4: Pomer neklasifikovaných kriviek v modele 2.



Obr. 5.6: Krabicové diagramy pre model 2.

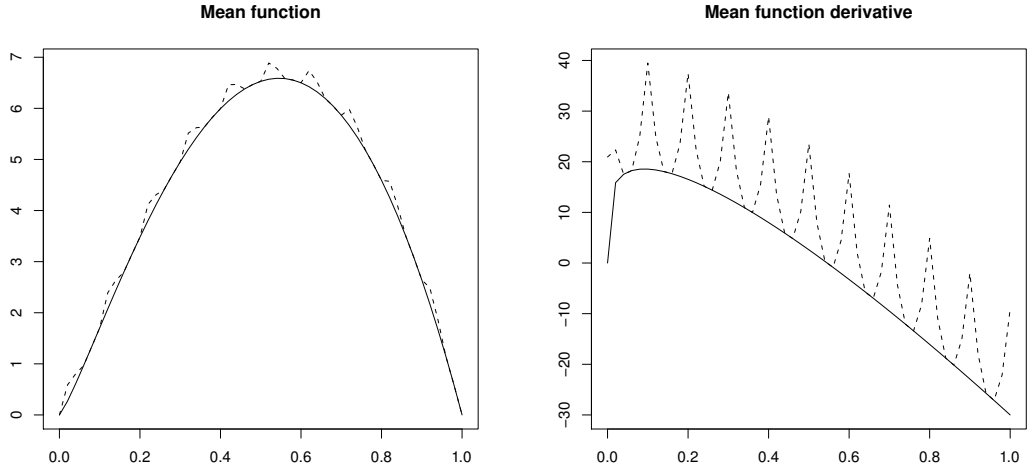
tribúcií má trajektórie je odlišného tvaru, zovšeobecnené pásové hĺbky nie sú schopné dostatočného rozoznania distribúcií.

Ak porovnáme 1-pásové hĺbkové funkcionály medzi sebou, vidíme výrazne lepšie výsledky pre voľbu obmedzenej 1-pásovej hĺbky podobne ako v modele 1.

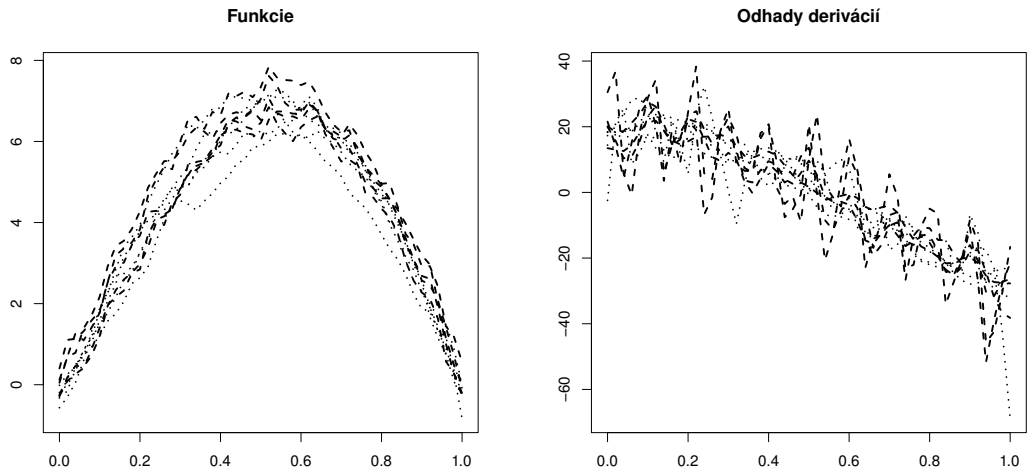
Pre účel zdôraznenia vhodnosti započítavania derivácií do funkcionálnej hĺbky dát urobíme ešte poslednú simuláciu s ďalším klasifikačným modelom posunutia v tvare. Funkcia strednej hodnoty $m_1(t)$ zostáva tvaru 5.3 a posunutá funkcia strednej hodnoty bude

$$m_2(t) = 30(1-t)t^{1.2} + \max\left(0, \frac{\sin(20\pi t)}{3}\right).$$

Funkcie stredných hodnôt a ich teoretické derivácie sú na obrázku 5.7, niekoľko realizácií na obrázku 5.8.

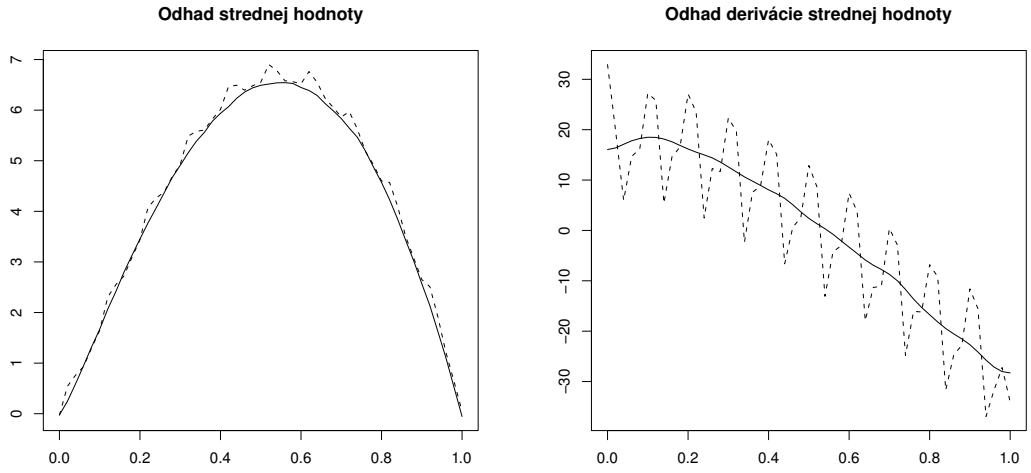


Obr. 5.7: Funkcie strednej hodnoty a ich prvé derivácie v modele 3.



Obr. 5.8: Niekoľko realizácií procesov v modele 3.

Poznamenajme, že pre funkciu strednej hodnoty $m_2(t)$ v modele 3 neexistuje obojstranná derivácia vo všetkých bodoch $t \in (0,1)$. Preto bude odhad derivácie tak ako bol uvedený v technických detailoch na začiatku kapitoly 5 chybný a výsledný odhad nebude podobný teoretickej derivácii na obrázku 5.7. Stredné hodnoty a ich prvé derivácie odhadnuté zo simulovaných dát sú na obrázku 5.9. Ako vidíme, odhad pr-



Obr. 5.9: Odhady funkcií strednej hodnoty a ich prvých derivácií v modele 3.

vej derivácie je viac podobný teoretickej prvej derivácii v modele 2 (obrázok 5.4) ako teoretickej derivácii v terajšom modele.

Dôvodom prečo sme teda vôbec uvažovali model 3 je, že aj v prípade, ak je odhad derivácie funkcií vysoko vychýlený, alebo v prípade ak derivácia sama o sebe neexistuje v niekoľkých bodoch definičného oboru funkcií, 1-pásová hĺbka môže poskytnúť oveľa lepšie rozoznanie tvaru funkcií stredných hodnôt ako všetky ostatné hĺbkové funkcionály. Výsledky tejto simulácie sú v tabuľkách 5.5 a 5.6 a krabicové diagramy

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0	0.02	0.52	0.6	0.67
1st Qu.	0	0.1	0.62	0.69	0.75
Median	0.01	0.16	0.65	0.73	0.78
Mean	0.0173	0.1531	0.6552	0.731	0.7838
3rd Qu.	0.03	0.1925	0.6925	0.77	0.82
Max.	0.08	0.3	0.8	0.86	0.91

Tabuľka 5.5: Pomer správne klasifikovaných kriviek v modele 3.

v obrázku 5.10.

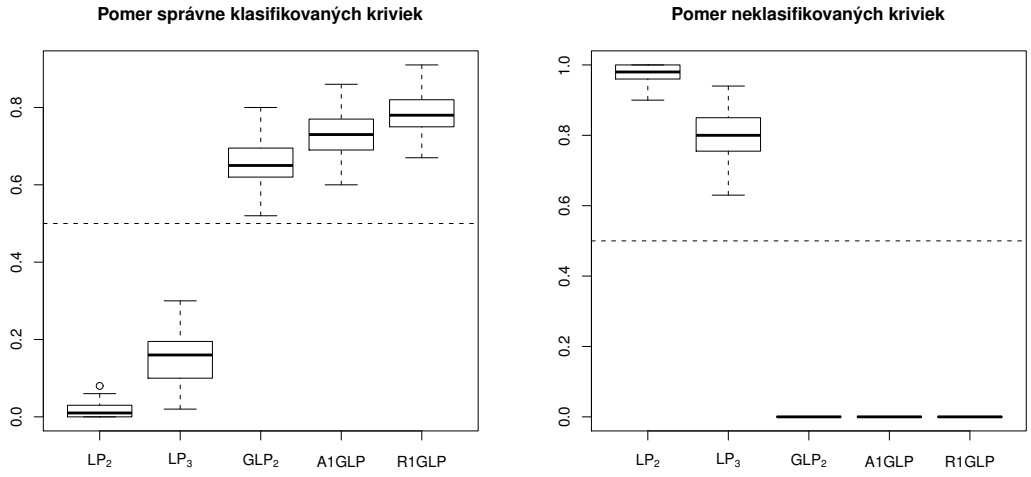
Rovnako ako v modele 2 poskytuje obmedzená 1-pásová hĺbka najlepšie odlíšenie funkcií z oboch simulovaných distribúcií.

5.3 Porovnanie na skutočných dátach - rast detí

Podobne ako Cuevas et al. [4] a López-Pintado a Romo [14] použijeme v poslednom príklade porovnania hĺbkových funkcionálov na základe klasifikácie známe dáta

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0.9	0.63	0	0	0
1st Qu.	0.96	0.7575	0	0	0
Median	0.98	0.8	0	0	0
Mean	0.9769	0.8045	0	0	0
3rd Qu.	1	0.85	0	0	0
Max.	1	0.94	0	0	0

Tabuľka 5.6: Pomer neklasifikovaných kriviek v modele 3.



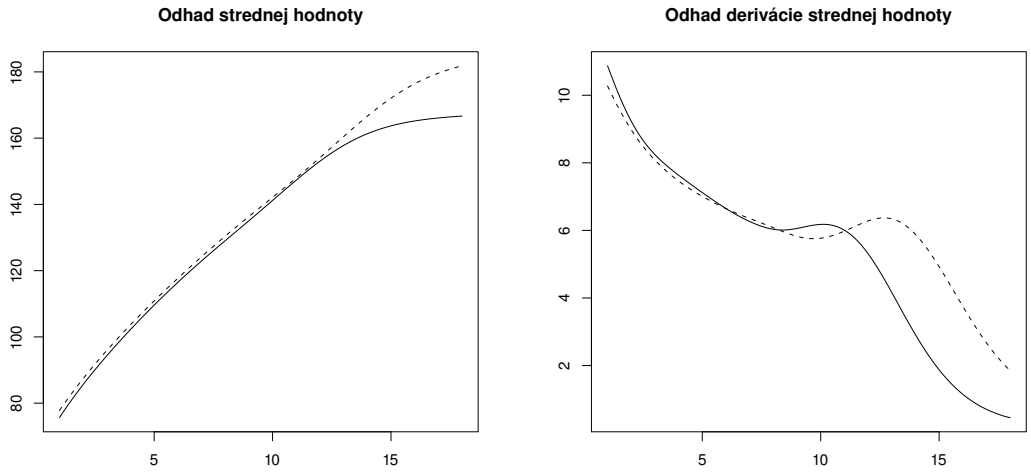
Obr. 5.10: Krabicové diagramy pre model 3.

detských rastových kriviek, ktorých časť sme už používali v príkladoch 9, 12 a 15. Tentokrát však využijeme kompletné dáta 54 dievčat a 39 chlapcov. V každom zo 100 nezávislých opakovaní vyberieme za tréningové skupiny náhodný podvýber s 30 pozorovaniami výšky chlapcov a 30 pozorovaniami výšky dievčat. Zvyšných 33 pozorovaní bude použitých ako krivky určené ku klasifikácii. Ostatné technické detaily sú rovnaké ako v prípade simulovaných dát z častí 5.1 a 5.2. Jediným rozdielom je, že funkčné hodnoty sú získané vyhladením metódou lokálnych kubických spline funkcií kvôli získaniu lepších vlastností týkajúcich sa monotónie odhadov. Odhad funkcií strednej hodnoty a ich prvých derivácií je na obrázku 5.11, niekoľko kriviek a ich derivácií na obrázku 5.12.

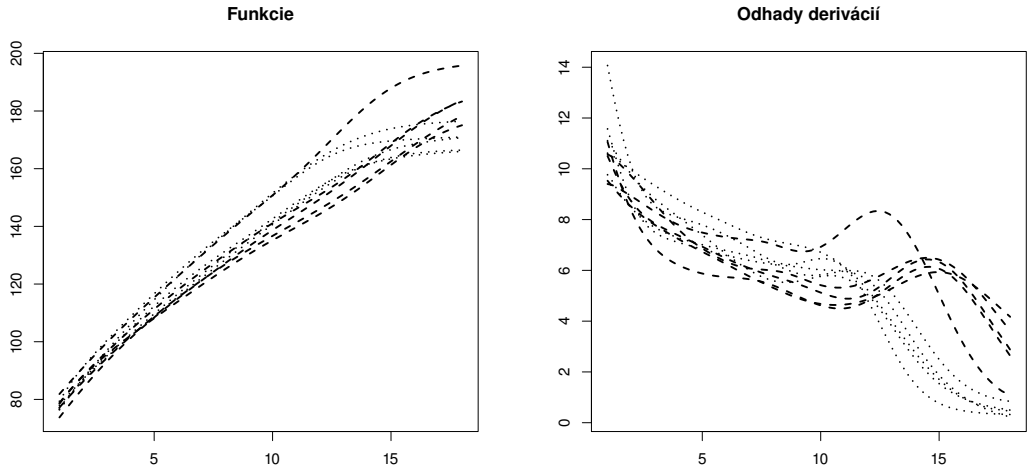
V dátach vidíme jasné posunutie v polohe funkcií strednej hodnoty od veku 15 rokov. Ďalej najmä z odhadov prvých derivácií vidíme posunutie v tvare funkcií od veku 10. To je spôsobené javom zvaným zrýchlenie rastu prítomným v puberte, pričom u dievčat k nemu dochádza o niečo skôr ako u chlapcov.

Výsledky klasifikácie našich dát môžeme vidieť v tabuľkách 5.7, 5.8 a na obrázku 5.13.

Čo sa týka interpretácie výsledkov, je jasné, že pásové hĺbky dávajú v tomto prípade porovnateľne dobré výsledky ako ostatné predstavené hĺbkové funkcionály. To je spôsobené tým, že v pozorovaniach bola prítomná veľmi malá zložka šumu. Preto pravdepodobnosť, že jedna funkcia v nejakom intervale leží mimo pásu tvoreného všetkými funkciami tréningovej skupiny nebola tak vysoká ako v simulovaných modeloch



Obr. 5.11: Odhady funkcií strednej hodnoty a ich prvých derivácií v modele rastu detí.



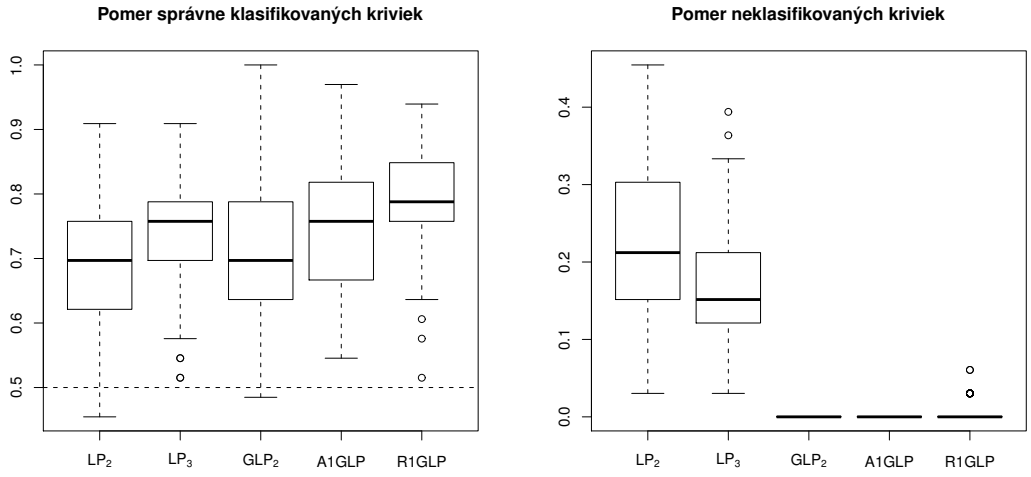
Obr. 5.12: Niekoľko kriviek v modele rastu detí.

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0.4545	0.5152	0.4848	0.5455	0.5152
1st Qu.	0.6288	0.697	0.6364	0.6667	0.7576
Median	0.697	0.7576	0.697	0.7576	0.7879
Mean	0.6864	0.7364	0.7121	0.7397	0.7927
3rd Qu.	0.7576	0.7879	0.7879	0.8182	0.8485
Max.	0.9091	0.9091	1	0.9697	0.9394

Tabuľka 5.7: Pomer správne klasifikovaných kriviek v modele rastu detí.

	$LP_n^{(2)}$	$LP_n^{(3)}$	$GLP_n^{(2)}$	$A1KLP_n$	$R1KLP_n$
Min.	0.0303	0.0303	0	0	0
1st Qu.	0.1515	0.1212	0	0	0
Median	0.2121	0.1515	0	0	0
Mean	0.2248	0.1645	0	0	0.006667
3rd Qu.	0.303	0.2121	0	0	0
Max.	0.4545	0.3939	0	0	0.06061

Tabuľka 5.8: Pomer neklasifikovaných kriviek v modele rastu detí.



Obr. 5.13: Krabicové diagramy pre model rastu detí.

s prítomnou šumovou zložkou. Napriek tomu pásové hĺbky a dokonca ani zovšeobecnená pásová hĺbka stále nedosahujú úspešnosť 1-pásových hĺbok, najmä obmedzenej 1-pásovej hĺbky.

Ak teda zhrnieme výsledky klasifikácie vo všetkých uvažovaných modeloch, klasifikácia na základe pásových hĺbok dáva katastrofálne výsledky v prípade zašumených pozorovaní. Ak sú ale pozorovania dostatočne hladké funkcie, úspešnosť pásových hĺbok rastie a môže byť porovnávaná aj so zovšeobecnenými pásovými hĺbkami. Na druhej strane si zovšeobecnené pásové hĺbky zachovávajú v oboch prípadoch dobrú spoľahlivosť a javia sa ako vhodné najmä v modeloch s posunutím v polohe. Môžu ale zlyhať v prípade rozoznávania funkcií odľahlých v tvare ak je odľahlosť v polohe nevýrazná alebo maskovaná blízkosťou funkcií.

Zo všetkých porovnaní boli jasne najlepšie výsledky dosiahnuté použitím obmedzenej 1-pásovej hĺbky. Táto zovšeobecnená pásová hĺbka, ktorá do výpočtu zahrnula derivácie dala dobré výsledky ako pre zašumené tak pre nezašumené pozorovania a bola schopná rozoznať rozdiely medzi distribúciami ako v polohe tak aj v tvare funkcií strednej hodnoty. Preto môžeme obmedzenú 1-pásovú hĺbku vyhlásiť za najlepší hĺbkový funkcionál z tých, ktoré sme v práci uvažovali.

Kapitola 6

Poznámky k výpočetnej náročnosti a záver

Na záver sa zamerajme na ďalšiu zaujímavú tému úzko spojenú s použiteľnosťou predstavených hĺbkových funkcionálov na riešenie praktických problémov. Vyšetrime, ako výpočetne náročné je vyhodnotenie hĺbky pre funkcionálne dáta. Vo všeobecnosti v prípade integrálnych hĺbok platí, že výpočetná zložitosť použitých metód je úmerná výpočetnej zložitosti jednorozmernej (alebo $(K+1)$-rozmernej v prípade K-pásových hĺbok) hĺbky funkčnej hodnoty v jednom bode definičného oboru voči príslušnému marginálnemu rozdeleniu.

V našich simuláciách v kapitole 5 sme použili integrálne hĺbky príslušné jednorozmernej ($GLP^{(2)}$) a dvojrozmernej simplexovej hĺbke ($R1KLP$ - obmedzená 1-pásová hĺbka), alebo ich konvexnej kombinácii ($A1KLP$ - priemerná 1-pásová hĺbka alebo všeobecne každá 1-pásová hĺbka s kladnými váhami). Na výpočet jednorozmernej simplexovej hĺbky je potrebných $O(n)$ operácií, pretože pre $x \in \mathbb{R}$ a $\mathbb{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ platí

$$SD_n(x; \mathbb{X}) = \binom{n}{2}^{-1} r(x; \mathbb{X}) (n - r(x; \mathbb{X})),$$

kde

$$r(x; \mathbb{X}) = \sum_{i=1}^{n} \mathbb{I}[x \leq X_i] \tag{6.1}$$

je počet pozorovaní s menšou hodnotou ako x.

Russeeuw a Ruts [22] našli rýchly algoritmus na výpočet dvojrozmernej simplexovej hĺbky založený na geometrických vlastnostiach simplexov. To je spôsob, akým je možné vylepšiť naivný algoritmus prehľadávania všetkých trojíc bodov náhodného výberu vyžadujúci $O(n^3)$ operácií na algoritmus vyžadujúci iba $O(n \log n)$ operácií. Ako poznamenali Russeeuw a Ruts [22], tento algoritmus je možné zovšeobecniť na výpočet d-rozmernej simplexovej hĺbky tak, aby jeho výpočetná náročnosť bola znížená na $O(n^{d-1} \log n)$ operácií. V našich výpočtoch sme použili takýto upravený algoritmus iba na 1-pásovú hĺbku, to znamená pre $d = 2$.

Výpočet pásovej hĺbky J-teho rádu vyžaduje $O\left(\binom{n}{J}\right) = O(n^J)$ operácií, pričom tento výraz s rastúcim počtom pozorovaní n veľmi rýchlo rastie najmä pre vyššie hodnoty J. Preto López-Pintado a Romo [15, 16] neodporúčajú používanie pásových hĺbok vyššieho rádu ako $J = 2$, $J = 3$ alebo v krajnom prípade $J = 4$. Výpočetná náročnosť sa však dá podstatne zlepšiť pomocou metódy podvýberov (resamplingu), ktorú navrhli López-Pintado a Jornsten [13]. Náhodný výber n funkcií je pri použití tejto metódy

náhodne rozdelený do $S \ll n$ skupín o približne rovnakej veľkosti. Výpočet pásovej hĺbky je potom prevedený voči každej z týchto skupín tak, ako by sa jednalo o samostatný náhodný výber. Priemer takýchto parciálnych hĺbok sa nakoniec považuje za aproximáciu pásovej hĺbky J-teho rádu funkcie voči celému náhodnému výberu. To, že takáto metóda môže dávať pomerne rozumné výsledky, ukázali López-Pintado a Jornsten [13]. Práve z dôvodu veľmi vysokej výpočetnej náročnosti pásovej hĺbky, vo všetkých simuláciách v kapitole 5 sme používali iba podvýberovú verziu pásových hĺbok s delením na skupiny približnej veľkosti 10.

Čo sa týka výpočetnej náročnosti zovšeobecnenej pásovej hĺbky, ukazuje sa situácia ako presný opak predchádzajúceho. Na výpočet zovšeobecnenej pásovej hĺbky každého rádu $J \in \mathbb{N}$, nie je potrebných viac ako $O(n)$ operácií. To platí, pretože j-ty sčítanec jednorozmernej konvexnej hĺbky $CD_n^j(.;\mathbb{X}) : \mathbb{R} \to [0,1]$ definovaný v 3.19 sa dá prepísať ako

$$\begin{aligned} CD_n^j(x;\mathbb{X}) &= \binom{n}{j}^{-1} \sum_{1 \le i_1 < \ldots < i_j \le n} \mathbb{I}\left[x \in \mathrm{Conv}\left(X_{i_1}, \ldots, X_{i_j}\right)\right] \\ &= \binom{n}{j}^{-1} \sum_{1 \le i_1 < \ldots < i_j \le n} \mathbb{I}\left[x \in \left[\min_{r=i_1,\ldots,i_j} X_r, \max_{r=i_1,\ldots,i_j} X_r\right]\right] \end{aligned}$$

a s využitím vlastností hypergeometrického rozdelenia sa dá vyjadriť ako funkcia $r(x) = r(x;\mathbb{X})$ zavedeného v 6.1

$$CD_n^j(x;\mathbb{X}) = 1 - \frac{\binom{r(x)}{j} + \binom{n-r(x)}{j}}{\binom{n}{j}}. \tag{6.2}$$

Navyše platí, že zovšeobecnená pásová hĺbka J-teho rádu $GLP_n^{(J)}$ sa dá počítať aj ako Fraimanov-Munizovej hĺbkový funkcionál indukovaný $CD_n^{(J)}$ (pozri 3.20). Ako sa dá ukázať s pomocou 6.2, existuje jednoduchý lineárny vzťah

$$CD_n^3(x;\mathbb{X}) = \frac{3}{2} CD_n^2(x;\mathbb{X}) \tag{6.3}$$

a preto ak označíme ako $X \in C^{(K)}([0,1])$ náhodnú funkciu, $\mathbb{X} = (X_1, \ldots, X_n)^T$ náhodný výber z $P \in \mathcal{P}\left(C^{(K)}([0,1])\right)$ a $\mathbb{X}(t)$ empirické marginálne rozdelenie funkčných hodnôt náhodného výberu $\mathbb{X}$ v bode $t \in [0,1]$, je

$$\begin{aligned} GLP_n^{(3)}(X;\mathbb{X}) &= \frac{1}{2}\left(\int_0^1 CD_n^2(X(t);\mathbb{X}(t))\,\mathrm{d}t + \int_0^1 CD_n^3(X(t);\mathbb{X}(t))\,\mathrm{d}t\right) \\ &= \frac{5}{4}\int_0^1 CD_n^2(X(t);\mathbb{X}(t))\,\mathrm{d}t \\ &= \frac{5}{4} GLP_n^{(2)}(X;\mathbb{X}). \end{aligned} \tag{6.4}$$

To znamená, že zovšeobecnená pásová hĺbka druhého a tretieho rádu sú si ekvivalentné.

Aj keď vzťahy medzi CD_n^2 a CD_n^j pre $j \ge 4$ sa už nedajú popísať lineárnou funkciou ako v 6.3 (a preto vzťah medzi $GLP_n^{(2)}$ a $GLP_n^{(J)}$ nie je možné vyjadriť jednoducho ako v 6.4), stále je možné zovšeobecnené pásové hĺbky J-teho rádu počítať iba pomocou $O(n)$ operácií pre každé J.

V tabuľke 6.1 môžeme jednoducho porovnať relatívny výpočetný čas potrebný k výpočtu hĺbok. Ľavý dolný index v značení pásových hĺbok znamená veľkosť podvýberov pri výpočte pomocou metódy podvýberov. Platí teda, že ${}_{10}LP_n^{(3)}$ znamená, že pásová hĺbka tretieho rádu bola počítaná pomocou metódy podvýberov do skupín o približnej veľkosti 10. Ak pri označení hĺbky ľavý dolný index nie je, znamená to, že hĺbka nebola počítaná metódou podvýberov.

n	GLP	$A1KLP_n$	${}_{10}LP_n^{(2)}$	${}_{10}LP_n^{(3)}$	${}_{10}LP_n^{(4)}$	$LP_n^{(2)}$	$LP_n^{(3)}$	$LP_n^{(4)}$
10	1	1.33	2.08	5.83	12.17	2.25	5.58	11.83
30	1.42	2.58	5.17	14.92	32.67	11.67	114.58	826.25
50	2.08	4.17	8	24.33	53.92	30	519.17	6505
100	2.33	7	15.33	47.5	105.17	125	4058.33	102625

Tabuľka 6.1: Relatívny výpočetný čas.

Ako vidíme, najrýchlejší výpočet je dosiahnutý pri použití zovšeobecnených pásových hĺbok, ale aj 1-pásové hĺbky dávajú stále veľmi dobré výsledky, dokonca aj pri pomerne veľkých rozsahoch náhodných výberov. Výpočetné časy v prípade pásových hĺbok však ukazujú ďalšiu obrovskú prekážku pri pokusoch o ich použitie na praktické účely. Pre vyšší rozsah náhodného výberu sa stáva prekážka pomalého výpočtu veľmi významnou už pre malé hodnoty rádov $J = 2$ alebo $J = 3$, v prípade vyššieho rádu $J = 4$ sa už presný výpočet pásovej hĺbky zdá byť takmer nedosiahnuteľný.

Výsledky pri porovnaní integrálnych hĺbok sú jasne viditeľné: najjednoduchšie Fraimanove-Munizovej funkcionály založené na simplexových hĺbkach sú jednoznačne lepšie ako pásové hĺbky. To platí z pohľadu rozoznávania vzorov vo funkcionálnych náhodných výberoch ako sme videli v kapitole 5, rovnako ako z pohľadu výpočetnej náročnosti. Jednorozmerné Fraimanove-Munizovej hĺbky však môžu zlyhať pri identifikácii skrytých funkcií odľahlých v tvare (funkcií odľahlých v tvare ale nie v polohe). Naopak K-pásové hĺbky dobre rozoznávajú ako funkcie odľahlé v tvare tak funkcie odľahlé v polohe. Na záver teda môžeme odporučiť používanie K-pásových hĺbok s vhodne volenými váhami vždy ak predpokladáme, že trajektórie náhodných procesov s ktorými pracujeme sú dostatočne hladké.

Literatúra

[1] Gérard Biau, Florentina Bunea, and Marten H. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory*, 51(6):2163–2172, 2005.

[2] Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.*, 10:340–355 (electronic), 2006.

[3] Probal Chaudhuri. Some intriguing properties of Tukey's half-space depth. *submitted to Bernoulli Journal*, 2010.

[4] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Comput. Statist.*, 22(3):481–496, 2007.

[5] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[6] Václav Dupač and Marie Hušková. *Pravděpodobnost a matematická statistika*. Nakladatelství Karolinum, Praha, 2005.

[7] Frank Falkner. *Child development:An international study*. Basel: Karger, 1960.

[8] Manuel Febrero, Pedro Galeano, and Wenceslao González-Manteiga. A functional analysis of NOx levels: location and scale estimation and outlier detection. *Comput. Statist.*, 22(3):411–427, 2007.

[9] Ricardo Fraiman and Jean Meloche. Multivariate L-estimation. *Test*, 8(2):255–317, 1999. With comments and a rejoinder by the authors.

[10] Ricardo Fraiman and Graciela Muniz. Trimmed means for functional data. *Test*, 10(2):419–440, 2001.

[11] Regina Y. Liu. On a notion of data depth based on random simplices. *Ann. Statist.*, 18(1):405–414, 1990.

[12] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist.*, 27(3):783–858, 1999. With discussion and a rejoinder by Liu and Singh.

[13] Sara López-Pintado and Rebecka Jornsten. Functional analysis via extensions of the band depth. In *Complex datasets and inverse problems*, volume 54 of *IMS Lecture Notes Monogr. Ser.*, pages 103–120. Inst. Math. Statist., Beachwood, OH, 2007.

[14] Sara López-Pintado and Juan Romo. Depth-based classification for functional data. In *Data depth: robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 103–119. Amer. Math. Soc., Providence, RI, 2006.

[15] Sara López-Pintado and Juan Romo. Depth-based inference for functional data. *Comput. Statist. Data Anal.*, 51(10):4957–4968, 2007.

[16] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. *J. Amer. Statist. Assoc.*, 104(486):718–734, 2009.

[17] Jaroslav Lukeš and Jan Malý. *Measure and integral*. Matfyzpress, Prague, second edition, 2005.

[18] Jean-Claude Masse and Jean-Francois Plante. *Depth: Depth functions tools for multivariate analysis*, 2009. R package version 1.0-1.

[19] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[20] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. Methods and case studies.

[21] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[22] Peter J. Rousseeuw and Ida Ruts. Bivariate location depth. *J. R. Stat. Soc., Ser. C*, 45(4):516–526, 1996.

[23] Robert Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons Inc., New York, 1980. Wiley Series in Probability and Mathematical Statistics.

[24] Robert Serfling. Depth functions in nonparametric multivariate inference. In *Data depth: robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 1–16. Amer. Math. Soc., Providence, RI, 2006.

[25] Robert Serfling. Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences, Second Edition*, 8:5338–5345, 2006.

[26] Charles J. Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(4):595–645, 1977. With discussion and a reply by the author.

[27] Read D. Tuddenham and Margaret M. Snyder. Physical growth of california boys and girls form birth to eighteen years. In *University of California Publications in Child Development*, pages 183–364, 1954.

[28] John W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congress, Montreal, Que., 1975.

[29] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.

Dodatok A

C++ zdrojové kódy

Základné C++ zdrojové kódy procedúr na výpočet hĺbok funkcionálnych dát (všetky zdrojové kódy všetkých použitých procedúr sú na priloženom CD, zdrojový súbor FunDepth.C a dynamická knižnica FunDepth.dll). Niekoľko detailov týkajúcich sa spôsobu výpočtu príslušných hĺbok:

Fraimanova-Munizovej polopriestorová hĺbka (zdrojový kód A.5)

Vo výpočte využívame vlastnosť polopriestorovej hĺbky v $\mathbb{R}$. V tomto prípade je polopriestorová hĺbka bodu $x \in \mathbb{R}$ voči náhodnému výberu $X_1, \ldots, X_n$ vyčísliteľná ako

$$HD_n(x; X_1, \ldots, X_n) = \min\left\{ \frac{\#\{X_i : X_i \leq x\}}{n}, 1 - \frac{\#\{X_i : X_i \leq x\}}{n} \right\}.$$

Pásová hĺbka (zdrojový kód A.6)

V skripte sa nepočíta celková pásová hĺbka funkcie x voči $\mathbb{X}$ o rozsahu n, výstupom je iba j-ty sčítanec $LP_n^j(x; \mathbb{X})$. Procedúra sa následne z R spúšťa $(J-1)$-krát pre všetky sčítance. Vstupom do nej sú okrem základných parametrov hlavne číslo komb, ktoré označuje počet kombinácií $\binom{n}{j}$ a com, ktorý obsahuje všetky tieto kombinácie. Napríklad všetky kombinácie 3 prvkov z 5 by vektor com obsahoval ako 123124125134135145234235245345.

Zovšeobecnená pásová hĺbka (zdrojový kód A.7)

Skript podobne ako pri pásovej hĺbke vyhodnocuje iba j-ty sčítanec hĺbky, výpočet je založený na 6.2.

1**-pásová hĺbka (zdrojový kód A.8)**

Základ vo výpočte 1-pásovej hĺbky tvorí skript preložený do jazyka C++ z originálneho Fortran skriptu výpočtu dvojrozmernej simplexovej hĺbky v balíku depth v programe R (naprogramovali Masse a Plante [18]). Táto procedúra cdepthint (nie je súčasťou prílohy, ale je súčasťou CD) pre bod v $\mathbb{R}^2$ vypočíta výberovú simplexovú hĺbku voči danému náhodnému výberu. Vo výpočte 1-pásovej hĺbky je kombinovaná s pomocnou funkciou djk (zdrojový kód A.4), ktorá počíta jednorozmernú konvexnú (teda špeciálne aj simplexovú) hĺbku.

To nám umožňuje pri voľbe parametrov $\alpha = (1,0)^T$ získať rýchlym výpočtom pomocou funkcie Kdepth aj Fraimanovu-Munizovej simplexovú hĺbku funkcie.

Výstupom zo skriptu je okrem 1-pásovej hĺbky funkcie aj dvojrozmerný vektor hlbfak, ktorý obsahuje príslušné sčítance 1-pásovej hĺbky a matica h, ktorá

obsahuje pre obe uvažované derivácie jedno- a dvojrozmernú simplexovú hĺbku v každom uvažovanom bode definičného oboru funkcií. Takto môžeme z matice h rekonštruovať funkcie, ktorých integrovaním cez definičný obor získavame sčítance 1-pásovej hĺbky

$$SD\left(x^{(0,\ldots,k)}(t), P_t^{(0,\ldots,k)}\right),\ t \in [0,1]$$

pre $k = 0, 1$.

Listing A.1: Hlavička.

```
#include <R.h>
#include <Rmath.h>
#include <stdlib.h>
#include <stdio.h>
// v každom C++ zdrojovom kóde platí:
// eval : počet bodov v diskretizácii funkcií
// m : počet funkcií v náhodnom výbere
// b : vektor o dĺžke eval, funkčné hodnoty funkcie ktorej hĺbku počítame
// b : pri 1−pásovej hĺbke matica funkčných hodnôt a prvých derivácií
// v : matica funkčných hodnôt m funkcií náhodného výberu v eval bodoch
// v : pri 1−pásovej hĺbke 3D−array funkčných hodnôt a prvých derivácií
// J : pri pásových hĺbkach označuje rád sčítanca, ktorý sa práve počíta
// kd : pri 1−pásovej hĺbke označuje K+1 (podporované len pre kd=1,2)
// alfa : pri 1−pásovej hĺbke označuje vektor váh o dĺžke kd
// Dj : pri pásových hĺbkach výsledný sčítanec pásovej hĺbky
// D : výsledná hĺbka funkcie
```

Listing A.2: Pomocná funkcia porovnávajúca dve čísla.

```
int compare(const void *x, const void *y) {
// x a y su pointre na double
// vracia −1 ak x < y
//    0 ak x == y
//    +1 ak x > y
        double dx, dy;
        dx = *(double *)x;
        dy = *(double *)y;
        if (dx < dy) {
    return −1;
                } else if (dx > dy) {
    return +1;
                }
        return 0;
        }
```

Listing A.3: Dvojica pomocných funkcií počítajúca kombinačné čísla.

```
int Kc(int M, int J){
// vracia kombinačné číslo M nad J ako int pre J=1,2,3,4
        if (M<J) return(0);
        else{
                if (J==1) return(M);
                if (J==2) return((M*(M-1))/2);
                if (J==3) return((M*(M-1)*(M-2))/6);
                if (J==4) return((M*(M-1)*(M-2)*(M-3))/24);
                }
        }

double Kc2(int M, int J){
// vracia kombinačné číslo M nad J ako double pre J=1,2,3,4
        if (M<J) return(0.0);
        else{
                if (J==1) return(M+0.0);
                if (J==2) return((M*(M-1))/2+0.0);
                if (J==3) return((M*(M-1)*(M-2))/6+0.0);
                if (J==4) return((M*(M-1)*(M-2)*(M-3))/24+0.0);
                }
        }
```

Listing A.4: Funkcia počítajúca jednorozmernú konvexnú hĺbku.

```
double djk(int r, int n,int j){
// vracia jednorozmernú konvexnú hĺbku pri danom r=r(x)
    return(1.0-(Kc2(n-r,j)+Kc2(r,j))/(Kc2(n,j)));
  }
```

Listing A.5: Fraimanova-Munizovej polopriestorová hĺbka.

```
void FTukey(int *eval, int *m, double *b, double v[*m][*eval], double *D){
  int i,j,k;
  *D = 0.0;
  for(i=0;i<*eval;i++){
    k = 0;
    for(j=0;j<*m;j++) if(v[j][i]<b[i]) k++;
    if ((*m-k)<k) k = *m-k;
    *D += (k+0.0)/(*m);
    }
  *D = *D/(*eval);
        }
```

Listing A.6: Pásová hĺbka.

```
void LP(int *eval, int *J, int *m, int *komb, int *com, double *b, double v[*m][*eval],
double *Dj){
// komb je koľko kombinácií prebehne
// com je vektor tvorený maticou combn
        double vpom[*eval][*J];
        int c,i,j;

        double LPindikator(int eval, int j, double *b, double v[eval][j]){
        // v je matica funkčných hodnôt j-tice funkcií z náhodného výberu,
        // pre ktoré práve LPindikator zisťuje, či vektor b leží v páse tvorenom nimi
                double mini, maxi,ind; // minimálny a maximálny element, indikátor
                int i,ii;
                ind = 1.0;
                i = 0;
                while ((i<eval) && (ind>0)){
                        mini = v[i][0];
                        maxi = v[i][0];
                        for (ii=0;ii<j;ii++){
                                if (mini>v[i][ii]) mini=v[i][ii];
                                if (maxi<v[i][ii]) maxi=v[i][ii];
                                }
                        if ((b[i]<mini) || (b[i]>maxi)) ind = 0.0;
                        i++;
                        }
                return(ind);
                }

        *Dj = 0.0;
        for (c=0;c<*komb;c++){
                for(i=0;i<*J;i++)for(j=0;j<*eval;j++)
                        vpom[j][i]=v[com[c*(*J)+i]-1][j];
                *Dj = *Dj + LPindikator(*eval,*J,b,vpom);
                }
        *Dj = *Dj/(*komb);
        }
```

Listing A.7: Zovšeobecnená pásová hĺbka.

```
void LPG(int *eval, int *J, int *m, double *b, double v[*m][*eval], double *Dj){
        int i,j,r;
        *Dj = 0.0;
        for(i=0;i<*eval;i++){
                r=0;
                for(j=0;j<*m;j++) if (v[j][i]<b[i]) r++;
                *Dj = *Dj + djk(r,*m,*J);
                }
        *Dj = *Dj/(*eval);
        }
```

Listing A.8: 1-pásová hĺbka.

```
1  void Kdepth(int *kd, int *eval, int *m, double b[*eval][*kd], double v[*m][*eval][*kd],
2  double *alfa, double h[*kd][*eval], double *D, double *hlbfak){
3          int i,j,k,l;
4          double U, V, X[*m], Y[*m];
5          for (i=0;i<*kd;i++) for (j=0;j<*eval;j++) h[i][j]=0.0;
6          for (j=0;j<*eval;j++){ // pre každý bod definičného oboru
7                  for (i=0;i<*kd;i++){ // pre každú deriváciu
8                          if (alfa[i]!=0){  // nech zbytočne neráta sčítance s váhou 0
9                                  U = b[j][0];
10                         // prvá súradnica bodu ktorý posudzujem (funkčná hodnota)
11                                 V = b[j][1];
12                         // druhá súradnica bodu ktorý posudzujem (hodnota derivácie)
13                                 for (k=0;k<*m;k++){
14                                         X[k] = v[k][j][0];
15                                 // prvé súradnice vektoru voči ktorému posudzujem
16                                         Y[k] = v[k][j][1];
17                                 // druhé súradnice vektoru voči ktorému posudzujem
18                                         }
19                                 if (i==0){   // funkčné hodnoty
20                                         l = 0;
21                                         for(k=0;k<*m;k++) if (X[k]<U) l++;
22                                         h[i][j] = djk(l,*m,2);
23                                         }
24                                 if (i==1) h[i][j] = cdepthint(U,V,*m,X,Y);
25                                 // derivácie
26                                 }
27                         }
28                 }
29         for (i=0;i<*kd;i++){
30                 hlbfak[i]=0.0;
31                 for(j=0;j<*eval;j++){ hlbfak[i]+=h[i][j]; }
32                 hlbfak[i]=hlbfak[i]/(*eval+0.0);
33                 }
34         *D = (alfa[0]*hlbfak[0]+alfa[1]*hlbfak[1])/(alfa[0]+alfa[1]);
35         }
```

Dodatok B

R implementácia zdrojových kódov

Implementácia C++ zdrojových kódov z prílohy A do programu R. Každý uvažovaný hĺbkový funkcionál je implementovaný v dvoch verziách: ako procedúra počítajúca hĺbku jednej funkcie voči náhodnému výberu (označené ako depth) a procedúra počítajúca pre náhodný výber hĺbku každej z funkcií voči ostatným funkciám náhodného výberu (označené ako sampledepth). Niekoľko písmen začínajúcich názov procedúry označuje hĺbku (F pre Fraimanove-Munizovej hĺbky, LP pre pásové hĺbky vrátane zovšeobecnených a K pre 1-pásovú hĺbku). Výstupom procedúry depth je štandardne jedna hodnota hĺbky funkcie voči náhodnému výberu (alebo v prípade komplikovanejšieho výstupu je celková hĺbka vo výstupe označená ako depth). Výstupom procedúry sampledepth je najmä vektor celkových hĺbok funkcií náhodného výberu sd. Všetky zdrojové kódy a podrobnejšia dokumentácia vstupov a výstupov jednotlivých funkcií sa dá nájsť na priloženom CD, skript ako súbor FunDepth.R a textový súbor dokumentácie ako FunDepth-dokumentacia.txt. Poznamenajme ešte, že indukované hĺbky funkcionálnych dát v kapitole 2.1 boli počítané pomocou funkcie PCdepth ktorá je tiež súčasťou FunDepth.R, tu sa ale nebudeme jej popisom zaoberať, pretože okrem indukovanej hĺbky počíta ešte ďalšie, v práci nepopísané funkcionálne hĺbky (viac dokumentácia k FunDepth).

Listing B.1: Hlavička.

```
library(depth)     # pre depth, nutne
dll = "FunDepth.dll" # nazov dll suboru so zdrojmi
# v každom R zdrojovom kóde platí:
# eval : počet bodov v diskretizácii funkcii
# m : počet funkcií v náhodnom výbere
# KD : pri 1−pásovej hĺbke označuje K+1 (podporované len pre kd=1,2)
# b : vektor o dĺžke eval, funkčné hodnoty funkcie ktorej hĺbku počítame
# b : pri 1−pásovej hĺbke matica funkčných hodnôt a prvých derivácií,
#   rozmery KD x eval
# v : matica funkčných hodnôt m funkcií náhodného výberu v eval bodoch,
#   rozmery eval x m
# v : pri 1−pásovej hĺbke 3D−array funkčných hodnôt a prvých derivácií,
#   rozmery KD x eval x m
# H : pri Fraimanovej−Munizovej hĺbke indikátor polopriestorovej verzie
#   TRUE=polopriestorová, FALSE=simplexová
# J : pri pásových hĺbkach označuje rád hĺbky (nie sčítanca ako v C++)
# G : indikátor generalizácie pásových hĺbok (TRUE pre zovšeobecnenú pásovú hĺbku)
# K : pri pásových hĺbkach počet resamplingových skupín na urýchlenie výpočtu
# alfa : pri 1−pásovej hĺbke označuje vektor váh o dĺžke KD
```

Listing B.2: Pomocná funkcia k výpočtu pásových hĺbok-jadro výpočtu.

```
LPKern = function(b,v,J=3,G){
# v ma rozmery evalxm
m = dim(v)[2]     # rozsah výberu
eval = length(b)   # eval
s = rep(0,J−1)    # medzivýpočet hĺbky
if (Load<−!is.loaded("LPindikator")) dyn.load(dll)
for (j in 2:J){    # az do J teho poradia
        if (G) s[j−1]=.C("LPG",as.integer(eval),as.integer(j),as.integer(m),as.double(b),
                as.double(v),dj = double(1))$dj
        if (!G){
                com = combn(m,j)
                poc = dim(com)[2]
                s[j−1]=.C("LP",as.integer(eval),as.integer(j),as.integer(m),
as.integer(poc),as.integer(com),as.double(b),as.double(v),dj = double(1))$dj
                }
        }
return(list(depth=mean(s),depthfactors=s))
}
```

Listing B.3: Pásové hĺbky:`depth`.

```
LPdepth = function(b,v,J=3,G,K=1){
# v ma rozmery evalxm
eval = length(b)
m = dim(v)[2]    # rozsah výberu
sam = sample.int(m,m)   # zamiešame
if (K==0) K=1    # opatrenie nech nemám K=0
nk = m%/%K    # velkost zhlukov
nlast = m %% K+nk   # velkost zostatkoveho zhluku
Dpom = matrix(rep(NA,K*(J-1)),nrow=K)
# medzivýpočet hĺbky, Dpom[k,j] je faktor hĺbky v k-tom zhluku j-tej hĺbky
if (Load<-!is.loaded("LPindikator")) dyn.load(dll)
if (K>1){
        for (k in 1:(K-1))
                if (eval>1) Dpom[k,] =
LPKern(b,v[,sam[((k-1)*nk+1):(k*nk)]],J,G)$depthfactors
                else Dpom[k,] =
LPKern(b,t(as.matrix(v[,sam[((k-1)*nk+1):(k*nk)]])),J,G)$depthfactors
        }
{if (nlast>0){
        if (eval>1) Dpom[K,]= LPKern(b,v[,sam[((K-1)*nk+1):m]],J,G)$depthfactors
                else Dpom[K,]=
LPKern(b,t(as.matrix(v[,sam[((K-1)*nk+1):m]])),J,G)$depthfactors
        }
        else (K=K-1)}
if (Load) dyn.unload(dll)
if(J>2) {if (K>1) fak = apply(Dpom[1:K,],2,mean) else fak = Dpom[1:K,]}
if(J==2) fak = Dpom[1:K]
return(list(depth=mean(Dpom[1:K,]),depthfactors=fak))
}
```

Listing B.4: Pásové hĺbky:`sampledepth`.

```
LPsampledepth = function(v,J=3,G,K=1){
m = dim(v)[2]
hlb = rep(NA,m)
hlbfak = matrix(rep(NA,m*(J-1)),nrow=(J-1))
dyn.load(dll)
for (i in 1:m){
        D = LPdepth(v[,i],v[,-i],J,G,K)
        hlb[i] = D$depth
        hlbfak[,i]= D$depthfactors
        }
dyn.unload(dll)
return(list(sd=hlb,sdfac=hlbfak))
}
```

Listing B.5: Fraimanove-Munizovej hĺbky:`depth`.

```
Fdepth = function(b,v,H){
if (H==FALSE) D = LPkern(b,v,J=2,TRUE)$depth
if (H==TRUE){
        if (Load<−!is.loaded("FTukey")) dyn.load(dll)
        D = .C("FTukey",as.integer(length(b)),as.integer(dim(v)[2]),as.double(b),
as.double(v),d=double(1))$d
        if (Load) dyn.unload(dll)
        }
return(D)
}
```

Listing B.6: Fraimanove-Munizovej hĺbky:`sampledepth`.

```
Fsampledepth = function(v,H){
if (H==FALSE) D = LPsampledepth(v,J=2,TRUE,K=1)$sd
if (H==TRUE){
        dyn.load(dll)
        m = dim(v)[2]
        D = rep(NA,m)
        for(i in 1:m) D[i] = Fdepth(v[,i],v[,−i],H)
        dyn.unload(dll)
        }
return(D)
}
```

Listing B.7: 1-pásová hĺbka:`depth`.

```
Kdepth = function(b,v,alfa=c(1,1)){
m = dim(v)[3]   # rozsah výberu, počet funkcií voči ktorým posudzujem
eval = dim(v)[2] # počet bodov v ktorých počítam
KD = dim(v)[1]
if (Load<−!is.loaded("Kdepth")) dyn.load(dll)
D = .C("Kdepth",kd=as.integer(KD),eval=as.integer(eval),m=as.integer(m),
b=as.double(b),v=as.double(v),alfa=as.double(alfa),h=double(eval∗KD),
hlbka=double(1),hlbfak=double(2))
if (Load) dyn.unload(dll)
return(list(depth=D$hlbka,depthfactors=D$hlbfak,pd=t(matrix(D$h,ncol=KD))))
# hlbfak su čiastočné hĺbky pre jednotlivé derivácie
}
```

Listing B.8: 1-pásová hĺbka:`sampledepth`.

```
Ksampledepth = function(v,alfa=c(1,1)){
# sample verzia
m = dim(v)[3]
D = rep(NA,m)
Df = matrix(rep(NA,dim(v)[1]*m),c(dim(v)[1],m)) # faktory hĺbky
spd = array(rep(NA,(dim(v)[1]*dim(v)[2]*dim(v)[3])),c(dim(v)[1],dim(v)[2],dim(v)[3]))
# sample point depth
# tretí rozmer určuje funkciu, ako v yn
dyn.load(dll)
for(i in 1:m) {
        kdpt = Kdepth(v[,,i],v[,,-i],alfa)
        D[i] = kdpt$depth
        #spd[,,i]=kdpt$pointdepth
        Df[,i]=kdpt$depthfactors
        }
dyn.unload(dll)
return(list(sd=D,sdfac=Df)) # chýba spd=spd
}
```